

**A MONTE CARLO COMPARISON OF POLYTOMOUS ITEM ESTIMATION BASED  
ON HIGHER-ORDER ITEM RESPONSE THEORY MODELS VERSUS HIGHER-  
ORDER CONFIRMATORY FACTOR ANALYSIS MODELS**

by

**Hong Wang**

B.A., Nan Kai University, 2002

M.A., University of Toledo, 2005

Submitted to the Graduate Faculty of  
School of Education in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Hong Wang

It was defended on

[September 14, 2012]

and approved by

Kevin H. Kim, Associate Professor, Psychology in Education

Clement A. Stone, Professor, Psychology in Education

Suzanne Lane, Professor, Psychology in Education

Levent Kirisci, Professor, Pharmaceutical Sciences

Dissertation Advisor: Kevin H. Kim, Associate Professor, Psychology in Education

Copyright © by Hong Wang

[2012]

# **A MONTE CARLO COMPARISON OF POLYTOMOUS ITEM ESTIMATION BASED ON HIGHER-ORDER ITEM RESPONSE THEORY MODELS VERSUS HIGHER-ORDER CONFIRMATORY FACTOR ANALYSIS MODELS**

Hong Wang, PhD

University of Pittsburgh, 2012

Item response theory (IRT) and confirmatory factor analysis (CFA) are two statistical techniques that were originally developed from different disciplines, but they are closely related to each other. Both can analyze the relationship between item responses and underlying constructs. This research investigated the performance of the two statistical methods with a higher-order structure in estimating polytomous response data. The higher-order IRT or second-order CFA model formulates correlational structure of multiple domains through a higher-order latent trait. This study compared Markov chain Monte Carlo (MCMC) estimation under a higher-order IRT model to mean-and-variance adjusted weighted least square (WLSMV) estimation under a second-order CFA model. The accuracy of the two estimation methods in recovering item parameters, overall and domain-specific abilities, and their correlations was examined under varied conditions.

The results showed MCMC and WLSMV methods were comparable on the accuracy of item discrimination and threshold parameter estimations. Although both estimation methods were found to yield more accurate item discrimination estimates as the number of items in each domain increased, WLSMV method was more sensitive to the number of items. The study also showed both estimation methods performed equally well in estimating overall and domain

abilities. The accuracy of ability estimation increased as the number of items, number of dimensions, and correlations between domains increased.

Beck Depression Inventory was analyzed using both estimation methods. Consistent with the findings in the simulation study, the results indicated the two estimation methods yielded quite comparable estimates for both item parameters and abilities at general and specific levels. The results also showed some variations in the item parameter estimates across different prior distributions used in MCMC, indicating the effect of priors on MCMC for the item parameter estimation. Furthermore, both estimation methods exhibited a convergence problem when the correlations between the general and specific factors were very high (i.e.,  $r > .90$ ).

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS .....</b>	<b>XII</b>
<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>2.0 LITERATURE REVIEW.....</b>	<b>8</b>
<b>2.1 UNIDIMENSIONAL ITEM RESPONSE THEORY.....</b>	<b>8</b>
<b>2.1.1 Unidimensional Dichotomous IRT Models .....</b>	<b>9</b>
<b>2.1.2 Unidimensional Polytomous IRT Models.....</b>	<b>11</b>
<b>2.2 MULTIDIMENSIONAL ITEM RESPONSE THEORY.....</b>	<b>16</b>
<b>2.2.1 Multidimensionality.....</b>	<b>16</b>
<b>2.2.2 Applications of MIRT .....</b>	<b>19</b>
<b>2.2.3 Hierarchical MIRT Models .....</b>	<b>22</b>
<b>2.3 COMPARISONS BETWEEN IRT AND CFA .....</b>	<b>26</b>
<b>2.3.1 Brief Introduction to CFA .....</b>	<b>26</b>
<b>2.3.2 Comparisons between IRT and CFA.....</b>	<b>30</b>
<b>2.4 ESTIMATION METHODS IN IRT AND CFA .....</b>	<b>34</b>
<b>2.4.1 Estimation Methods in IRT .....</b>	<b>34</b>
<b>2.4.1.1 Marginal Maximum Likelihood (MML) vs. Unweighted Least Square (ULS).....</b>	<b>34</b>

2.4.1.2	Markov chain Monte Carlo (MCMC) Simulation .....	36
2.4.2	Standard Estimation Methods in CFA .....	46
2.4.3	Previous Studies on the Comparisons of Estimation Methods across IRT and CFA .....	48
3.0	METHOD .....	52
3.1	MODEL SPECIFICATION .....	52
3.1.1	Higher-order Graded Response IRT Model .....	52
3.1.2	Second-order CFA Model .....	54
3.2	SIMULATION STUDY .....	56
3.2.1	Study Design .....	56
3.2.2	Simulation Procedure .....	59
3.2.3	Validation of Data Generation .....	63
3.3	ESTIMATION .....	65
3.4	EVALUATION CRITERIA .....	73
4.0	RESULTS .....	76
4.1	ITEM PARAMETER RECOVERY .....	76
4.2	RECOVERY OF REGRESSION COEFFICIENTS .....	80
4.3	ABILITY RECOVERY .....	83
4.4	REAL DATA APPLICATION .....	86
5.0	DISCUSSION .....	92
5.1	SUMMARY AND IMPLICATIONS .....	92
5.1.1	MCMC vs. WLSMV in Parameter Estimation .....	92
5.1.2	Effects of Between-subjects Independent Variables .....	93

5.1.3	Implications for Practice.....	96
5.2	LIMITATIONS AND FUTURE RESEARCH DIRECTIONS .....	98
	APPENDIX A .....	100
	APPENDIX B .....	104
	APPENDIX C .....	106
	APPENDIX D .....	107
	BIBLIOGRAPHY .....	108



## LIST OF TABLES

Table 3.1 Simulated IVs and Levels for Each IV .....	56
Table 3.2 An Example Set of Simulated Item Parameters.....	61
Table 3.3 Observed and Expected Proportions of Response Categories .....	63
Table 3.4 Factor Pattern for Exploratory Factor Analysis with Promax Rotation.....	64
Table 3.5 Recovery of Regression Coefficients of Domain Abilities on Overall Ability .....	65
Table 3.6 True and Estimated Parameter Values in WinBUGS .....	70
Table 4.1 Mean and SE of Item Discrimination RMSE by Between-subjects IVs .....	77
Table 4.2 RMSE and BIAS of Item Discrimination Estimated by MCMC vs. WLSMV .....	78
Table 4.3 RMSE and BIAS of Item Threshold Parameters Estimated by MCMC vs. WLSMV .	80
Table 4.4 Mean and SE of RMSE of Regression Coefficients by Between-subjects IVs .....	81
Table 4.5 RMSE of Regression Coefficients Estimated by MCMC vs. WLSMV .....	82
Table 4.6 Mean and SE of Correlation between Estimated and True Overall Ability by Between-subjects IVs.....	84
Table 4.7 Correlation between True and Estimated Overall Ability .....	84
Table 4.8 Correlation between True and Estimated Domain Abilities for D=3 .....	85
Table 4.9 Correlation between True and Estimated Domain Abilities for D=5 .....	86
Table 4.10 Mean Item Parameter Estimates for the BDI Subscales .....	89

Table 4.11 Estimated Regression Coefficients between the Overall and Subscale Scores for the BDI Data .....	90
---	----

Table 4.12 Correlation of Ability Estimates between MCMC and WLSMV for the BDI Data...	91
--	----

## LIST OF FIGURES

Figure 2.1 Item characteristic curve for an item with $a=1$ , $b=0.5$ , and $c=0.1$ .....	10
Figure 2.2 Operating characteristic curves for a five-category item under the GR model.....	14
Figure 2.3 Category response curves for a five-category item under the GR model.....	16
Figure 2.4 Two types of multidimensionality with two dimensions and six items .....	18
Figure 2.5 Illustration of a HO-IRT model.....	23
Figure 2.6 Example of a bi-factor model .....	30
Figure 3.1 Subscale information for a three-dimension test with five items in each dimension..	62
Figure 3.2 History plots of discrimination and threshold parameters for item 1 .....	67
Figure 3.3 BGR diagnostic plots of discrimination and threshold parameters for item 1 .....	68
Figure 3.4 Autocorrelation plots of discrimination and threshold parameters for item 1.....	69
Figure 3.5 History plots of $\gamma$ estimates for three domain abilities .....	71
Figure 3.6 Autocorrelation plots of $\gamma$ estimates for three domain abilities.....	72
Figure 3.7 BRG plots of $\gamma$ estimates for three domain abilities.....	72
Figure 4.1 RMSE of Item discriminations by estimation methods and number of items.....	78
Figure 4.2 RMSE of regression coefficients by estimation method and correlation levels between domains .....	82

## **ACKNOWLEDGEMENTS**

I would like to take this opportunity to express my sincere gratitude to all who have encouraged and supported me to complete this dissertation and my doctoral study.

First, I would like to extend my greatest appreciation to my advisor, Dr. Kevin Kim, for his expert guidance and constant support throughout my doctoral study. I am thankful to Kevin, particularly for his prompt and insightful feedback to all my questions and great patience at every stage of my dissertation research. It was due to his paramount mentorship that I was able to complete this project.

I would also like to thank three other members in my dissertation committee: Dr. Clement Stone, Dr. Suzanne Lane, and Dr. Levent Kirisci for their constructive and valuable input for improving the quality of my research. My special thanks also go to all the faculty members in the Research Methodology program who helped me to grow rapidly as a fledging scholar. I am thankful to Dr. Clement Stone, under whose guidance I built a solid foundation in educational measurement and computer programming. I would like to thank Dr. Suzanne Lane for her great encouragement and a variety of opportunities she provided for my professional development. I am also grateful to Dr. Feifei Ye for her excellent instruction and guidance in developing my knowledge and research skills in statistics. In addition, I appreciate all the

wonderful get-together parties she organized which undoubtedly added enormous fun to my life in Pittsburgh.

Finally, I would dedicate my deepest gratitude to my family members. I would like to thank my parents and my sister who always encourage and support me with their enduring love. My special love and thanks also go to my husband, Jing Xu, who always believes in me and encourages me to complete the hard work. I'm indebted to his endless love, caring, and support.

## **1.0 INTRODUCTION**

Item response theory (IRT) and confirmatory factor analysis (CFA) are two types of measurement techniques that were originally developed from different disciplines, but they are closely related to each other. Both classes of models can be used to analyze the relationship between item responses and underlying constructs. Particularly, both are applicable to item-level data that are ordered categories: either dichotomous (e.g., multiple-choice items with correct vs. incorrect responses) or polytomous (e.g., an attitude item with responses strongly disagree, disagree, agree, and strongly agree). The comparison between the two techniques has been an important topic as both frameworks provide the means of evaluating the psychometric properties of measurement instruments. The current study intends to further assess the performance of the two techniques under a more complex model structure.

IRT is a family of statistical models that relate observed item responses to respondents' latent trait levels. The application of IRT is commonly found in education, where it has been used for developing and refining tests, scoring examinees, and equating different test forms. Meanwhile, IRT has also been applied to a wide range of psychological areas such as intelligence tests, personality assessments, and clinical measures (Embretson & Reise, 2001; Reise & Waller, 2009). Traditional IRT assumes unidimensionality, that is, all the items in a test measure a single common construct. However, there are many testing situations that require the measurement of multiple content domains or subscales. In the field of education, for instance, the

ACT test is broken down to multiple sections to measure different content domains (English, reading, math, and science). In the field of psychology, there are also many tests that include multiple subscales. For example, the BDI-II (Beck Depression Inventory-II), a widely used instrument for measuring the severity of depression, includes two subscales: affective component and physical component. The purpose of subscales is to help identify respondents' strengths or weaknesses (like ACT) or to determine the primary cause of respondents' behavior (like BDI-II). For a scale with multiple subscales, unidimensional IRT that treats all the subscales as one construct might not be appropriate as the distinctions among different subscales are ignored. Many studies on the consequences of applying unidimensional IRT to multidimensional data have shown that the resulting parameter estimates may be biased (e.g., Ackerman, 1989; Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988). In order to differentiate a variety of subscales, it is also possible to use unidimensional IRT for each subscale separately. However, results derived from such an approach are usually unreliable or less accurate due to the small number of items included in each subscale. Moreover, subscales are usually correlated, and the correlation among subscales would be ignored if unidimensional IRT was used. Therefore, between-item multidimensional item response theory (MIRT) has been proposed to simultaneously analyze response data from a test involving multiple subscales (e.g. de la Torre, 2008; de la Torre & Patz, 2005; Sheng & Wikle, 2007).

MIRT explicitly models a scale with multiple dimensions and accounts for possible correlations among dimensions. Many studies have shown that MIRT provided more accurate and efficient estimation on subscales than unidimensional IRT. It was also more efficient with multiple short tests that were highly correlated (de la Torre, 2008, 2009; de la Torre & Patz, 2005; Hong, Lam, & de la Torre, 2010; Sheng & Wikle, 2008; Yao & Boughton, 2007). An

alternative to MIRT is higher-order IRT (HO-IRT) (de la Torre & Song, 2009). For HO-IRT, in addition to subscales (often referred to as domain abilities), an overall ability is posited above these domain abilities. A linear relationship is assumed between the overall ability and domain abilities. The correlations among domain abilities are determined by these linear relationships. One advantage of HO-IRT over MIRT is that the respondents' performance on subscales and their overall performance can be estimated simultaneously. This method satisfies the requirement of reporting both overall and domain scores in many testing situations. For example, for an English proficiency test with three subtests: listening, reading, and writing, one may be interested in the overall scores to indicate a general proficiency in English as well as scores on each subscale to provide some diagnostic information for a future improvement. Although unidimensional IRT can estimate an overall ability, this method may not be valid because of the extent to which the unidimensional assumption is violated (de la Torre, 2009). Hence, HO-IRT that includes both overall and specific abilities in one model is more appropriate.

Developed almost in parallel with IRT, CFA is another class of statistical models used to measure the relationship between observed variables and constructs (often called factors in CFA). CFA is the core measurement component of structural equation modeling (SEM). Although developed independently, IRT and CFA are closely related to each other. The multiple dimensions or subscales in MIRT correspond to factors in CFA. The latent structure indicated by HO-IRT is, in fact, equivalent to second-order CFA. A second-order CFA is more frequently used for testing a theory. For example, in intelligence research, one of popular intelligence theories states that the more specialized facets of ability (e.g., verbal comprehension, perceptual organization, memory) are influenced by a general intelligence (g) (Brown, 2006). Both HO-IRT and second-order CFA conform to such a latent structure. In fact, the analytical relationship



between IRT and CFA has been established more than two decades ago by Takane and de Leeuw (1987). They mathematically demonstrated the equivalence of IRT and CFA for dichotomous response data. The similarities and differences between the two techniques have been discussed both theoretically and/or empirically (e.g., Glöckner-Rist & Hoijtink, 2003; Reckase, 1997a; Tate, 2003; Wirth & Edwards, 2007). Each methodology has its own strengths and weaknesses. For example, Glöckner-Rist and Hoijtink (2003) pointed out that CFA is relatively weak in making inferences at the person level (latent traits), whereas IRT is less flexible in the integration of measurement components and structural model which measures the relationship between factors and other manifest variables.

Another practical distinction between IRT and CFA is the estimation methods used by the two techniques. In the framework of IRT, particularly in MIRT, the comparisons among estimation methods were usually accompanied with the estimation software such as TESTFACT (Bock, et al., 2002) and NOHARM (Normal-Ogive Harmonic Analysis Robust Method, Fraser & McDonald, 1988). Although both programs were developed under the IRT framework, the estimation methods were, in fact, heavily borrowed from factor analysis and SEM approaches (Reckase, 1997a). One notable limitation for both programs is that neither can be used to analyze multidimensional polytomous item responses. With the rise of Markov chain Monte Carlo (MCMC) algorithm (Patz & Junker, 1999a, 1999b), more complex models such as multidimensional polytomous response data can be handled (e.g. Edwards, 2010; Yao, 2010). MCMC is becoming a popular estimation technique for MIRT. Such complex models can also be estimated using CFA software such as Mplus (Muthén & Muthén, 1998-2010), EQS (Bentler, 1995), and LISREL (Jöreskog & Sörbom, 1996). For these computer programs, the most commonly used estimation method of ordered-category response data is weighted least square

(WLS), or variations of WLS such as mean adjusted WLS (WLSM) and mean-and-variance adjusted WLS (WLSMV, Muthén, du Toit, & Spisic, 1997). Previous studies have demonstrated that MIRT and CFA produced comparable item parameter estimates for dichotomous response data (Finch, 2010; Knol & Berger, 1991; Tate, 2003; Wirth & Edwards, 2007).

However, the comparisons across the two techniques have been limited to first-order factor models. The similarities or differences between the higher-order IRT and CFA have not been examined. Meanwhile, few of prior studies have examined the performance of the two methods in the analysis of polytomous responses, which are commonly used in performance-based educational assessments or questionnaires for measuring various psychological constructs. Moreover, the investigation of HO-IRT has been mainly focused on dichotomous response data (e.g., de la Torre & Hong, 2010; de la Torre & Song, 2009). Hence, the primary purposes of the present study are 1) to extend HO-IRT to polytomous response data, and 2) to compare the performance of HO-IRT and second-order CFA in recovering item and person parameters with polytomous response data. Specifically, HO-IRT with MCMC was compared to second-order CFA with WLSMV. The following major research questions were addressed:

- 1) Is there any difference between the two methods in the recovery of the item parameters, i.e., difficulty and discrimination parameters?
- 2) Do the two methods perform equally in estimating regression coefficients of domain abilities on the overall ability, i.e., recovery of correlation between dimensions?
- 3) Do the two models perform equally in estimating overall ability/second-order factor scores?
- 4) Is there any difference between the two models in the estimation of domain abilities/first-order factor scores?

In order to answer these research questions, a simulation study was conducted. The goal of this study, first of all, is to add to the existing body of literature on the comparison between CFA and IRT by investigating the parameter estimations of models with a higher-order factor. Secondly, by extending the application of HO-IRT to polytomous items, this study intends to compare its performance to second-order CFA for polytomous response data. Most previous studies on the comparisons between IRT and CFA have been focused on dichotomous responses (e.g. Finch, 2010; Knol & Berger, 1991; Tate, 2003). Polytomous items, however, have been commonly applied in many measurement instruments. In the field such as educational measurement, the use of constructed response items or other forms of performance assessment have gained popularity. These performance-based assessments attempt to measure how the intended knowledge and skills are applied or approached by the examinees in the real-world-like context. Unlike multiple-choice items that are usually scored dichotomously (i.e., correct vs. incorrect), the responses constructed by examinees in a performance assessment are usually scored on a rubric with more than two levels. The performance-based assessments have been increasingly used in many large-scale assessments and accountability programs (Lane & Stone, 2006). For example, in the SMARTER Balanced Assessment Consortium (SBAC) system, a recently established assessment system in response to the Race to the Top program, the proposed summative assessments are composed of not only an adaptive test including a variety of item formats but also solely performance tasks in two content domains (English languages arts and mathematics). In addition, polytomous items are quite commonly seen in many psychological measures. Particularly, many of these psychological instruments include multiple dimensions or subscales. For instance, Hill et al. (2007) used IRT to analyze a pediatric quality of life inventory which includes four subscales with each measured by 5 or 7 items scored on a 5-point response

scale. Motivated by the application of polytomous items and multidimensional model structure indicated by many measurement instruments, this study attempts to extend the comparison between HO-IRT and second-order CFA to polytomous response data.

Another important issue addressed by the current study is to compare the estimation methods used in the two frameworks. As a promising estimation method in IRT, MCMC has seldom been compared to popular CFA estimation methods. Hence, this study focuses on comparing the performance of the MCMC and WLSMV estimation methods for higher-order IRT and second-order CFA. Finally, the study seeks to investigate how different testing conditions such as the number of dimensions, test length, sample size, and correlation between dimensions affect parameter estimates. Examining the performance of the two methodologies across a wide range of conditions will be helpful for identifying the strengths and weaknesses of each methodology, which can further guide applied researchers to choose from or to integrate the two methodologies.

The remainder of this dissertation is organized as follows. The second chapter reviews related literature on IRT and CFA to provide some theoretical background. The third chapter presents the research design, simulation procedure, and evaluation criteria used for this study. The fourth chapter describes the results from both simulated and real data. Finally, the fifth chapter provides a discussion of the results as well as the limitations and future directions of the current research.

## **2.0 LITERATURE REVIEW**

The purpose of this project is to examine the performance of two methods (IRT with MCMC vs. CFA with WLSMV) for multidimensional polytomous data with a higher-order factor. To provide the necessary background information, this chapter presents related literature including 1) unidimensional IRT, 2) multidimensional IRT, 3) comparisons between IRT and CFA, and 4) estimation methods commonly used in IRT and CFA.

### **2.1 UNIDIMENSIONAL ITEM RESPONSE THEORY**

IRT is a family of statistical models that are commonly used in psychological and educational measurement to relate respondents' observed behavior to their latent trait/ability levels. According to different criteria, IRT models can be classified in various ways. For example, based on the number of traits measured, IRT models can be generally divided into two families: unidimensional and multidimensional. Unidimensional models assume a single trait accounting for examinees' performance. Multidimensional IRT models, on the other hand, model examinee's performance in relation to multiple traits. Another example of classifying IRT models is based on the number of scored responses. Dichotomous IRT models are applied to test items that have two categories such as multiple-choice items (correct vs. incorrect). Polytomous IRT models are used for items that have more than two categories such as open-ended

mathematics problems (scoring on a rubric from 0 to 3) or rating scales for survey questions (rating on a scale from 1 to 5). This section mainly reviews some important unidimensional IRT models.

### 2.1.1 Unidimensional Dichotomous IRT Models

In a dichotomous IRT model, the probability of examinee's correct response to an item (usually scored 0 or 1) is a mathematical function of both person's trait and properties of the items. Two commonly known mathematical functions used in IRT models are logistic and normal ogive. Both functions, in fact, are found to produce very similar probabilities and parameter estimates, but logistic models are more often used due to their computation simplicity (Embretson & Reise, 2000). In addition to the mathematical functions, IRT models can be further classified according to the number of item parameters involved. For instance, for a 3-parameter logistic model (3PL), the probability of a correct response to an item  $i$  is

$$p(x_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp^{-Da_i(\theta_j - b_i)}} \quad (2.1)$$

where

$x_{ij}$  is the response of person  $j$  to item  $i$  (scored 0 or 1, and 1 presents correct response);

$\theta_j$  is the person parameter (latent trait);

$a_i$  is the discrimination of item  $i$ ;

$b_i$  is the location (or 'difficulty' in educational measurement) of item  $i$ ;

$c_i$  is the lower asymptote (guessing) of item  $i$ ;

$D$  is the scaling constant (1 or 1.7).

This model defines the probability of a correct response to an item as a function of one person parameter ( $\theta_j$ ) and three item parameters ( $a_i$ ,  $b_i$ , and  $c_i$ ). The relationship among the item parameters can be depicted by an item characteristic curve (ICC, Figure 2.1). The item difficulty parameter  $b_i$  corresponds to the item location in the ICC. Note that the model scales the item's difficulty and person's trait  $\theta$  onto the same continuum. Therefore, the item difficulty is the point on an ability scale where the ICC has its maximum slope. The item discrimination parameter  $a_i$  represents the slope of the ICC which describes how rapidly the probabilities change along the trait level. The steeper the slope, the greater extent to which the item discriminates between persons. The guessing parameter  $c_i$  is used to account for the effects of guessing on the probability of a correct response, which often occurs in multiple-choice items. Represented as a lower asymptote in the ICC, it indicates the probability that very low ability individuals will get this item correct by chance.  $D$  is a scaling constant.  $D = 1.7$  is typically used to make the item parameters from the logistic IRT model very similar to the item parameters that would be obtained in the normal-ogive IRT model.

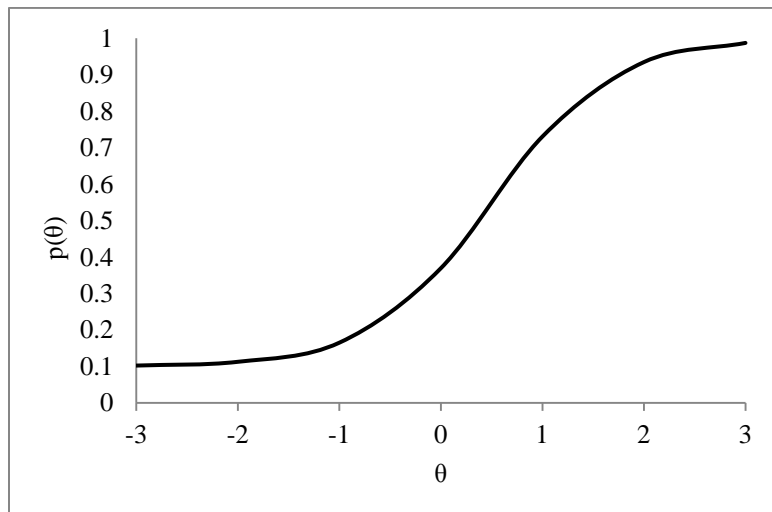


Figure 2.1 Item characteristic curve for an item with  $a=1$ ,  $b=0.5$ , and  $c=0.1$

When the guessing parameter  $c_i = 0$  in Equation (2.1), the resulting model corresponds to two-parameter logistic model (2PL). This model is appropriate for testing items where there is no guessing. For example, guessing seems irrelevant for the item "I like coke better than pepsi" with responses of agree/disagree. There is another class of models called one-parameter logistic (1PL) or Rasch model where only one parameter, the item difficulty ( $b_i$ ), is needed to describe the response data. In other words, Rasch model assumes all items have the same slope.

### **2.1.2 Unidimensional Polytomous IRT Models**

Polytomous IRT models are commonly used for measurement instruments that include items with multiple categories. In polytomous IRT models, the probability of responding in a particular category is a function of the examinee's trait level and item properties. There are various unidimensional polytomous models available, and the most common ones include graded response (GR; Samejima, 1969), partial credit (PC; Masters, 1982), generalized partial credit (GPC; Muraki, 1992), rating scale (RS; Andrich, 1978), and nominal response (NR) models (Bock, 1972). According to Thissen and Steinberg's (1986) taxonomy for item response models, these polytomous response models can be categorized into 'difference' model and 'divide-by-total' model. Samejima's graded response model falls under the rubric of 'difference' model where the probability of responding in a particular category is obtained by taking the difference between the probability of scoring at or above level  $k$  and the probability of scoring at or above level  $k+1$ . The remaining models are 'divide-by-total' models. For this type of model, the probability of responding in a particular score level can be expressed directly as the ratio of the function for that level to the sum of the functions for all the levels (Yen & Fitzpatrick, 2006).



NR model is the most general divided-by-total model because it incorporates nominal responses, i.e., the response categories are not ordered. All other divide-by-total models can be regarded as special cases of NR model. The selection of models primarily depends on the type of item data. For example, if the items are assumed to have equal discrimination, then PC and RS models might be sufficient because these two polytomous models are extensions of Rasch model. Nevertheless, RS model cannot be applied to a set of items that differ in the number of levels. For example, if a test includes a combination of items with 3-point and 4-point scales, then RS model is not applicable because it assumes a fixed set of rating scale used for the entire test. In contrast, when the items are suspected to have different discriminations, GR or GPC model, both of which include discrimination parameters, is naturally more appropriate than any Rasch-based models. However, GR model is distinguished from GPC model by the fact that it requires a two-stage process to compute the conditional probability for an examinee responding to a particular category. Furthermore, GR model has an order restriction on category locations in comparison to GPC model. GR model requires that category locations (often called threshold parameters between levels) should be ordered within each item. This is, however, not required by GPC model which stresses the relative difficulty of each step of transition from one category to the next in an item. These steps in GPC model are allowed to be relatively easier or more difficult than others within an item.

Samejima's GR model was used in this study and therefore explained in more detail. GR model can be viewed as a generalization of the dichotomous 2PL model. It is appropriate for items with ordered polytomous responses such as Likert or constructed response items. In GR model, two steps are implemented to compute category response probabilities. The first step is to calculate cumulative category response functions that represent the probability of scoring at or

above particular category level  $x$  ( $x = 0, 1, \dots, m_i$ ) on an item. What essentially occurs in this step is that the response of each item is dichotomized into two overall categories: (1) greater or equal to category level  $x$  and (2) less than category level  $x$ . For example, for an item with five-response options  $x = 0, 1, \dots, 4$ , GR model treats the item responses as four dichotomies, i.e., (0 vs. 1 to 4); (0 and 1, vs. 2 to 4); (0 to 2 vs. 3 and 4); and (0 to 3 vs. 4). Then, the probability ( $P_{ijx}^*$ ) for examinee  $j$  for category  $x$  ( $x = 0, \dots, m_i$ ) or higher on item  $i$  can be modeled using the 2PL function

$$P_{ijx}^* = \frac{\exp[Da_i(\theta_j - b_{ix})]}{1 + \exp[Da_i(\theta_j - b_{ix})]} \quad (2.2)$$

where

$\theta_j$  is the ability for examinees  $j$ ;

$a_i$  is the discrimination or slope parameter of item  $i$ ;

$b_{ix}$  is the threshold parameter for category  $x$  of item  $i$ ;

$D$  is the scaling constant (1 or 1.7).

The  $b_{ix}$  parameters represent the boundaries between category levels. For an item with  $(m_i + 1)$  response categories, there are  $m_i$  threshold parameters and one item discrimination parameter ( $a_i$ ). For each threshold parameter, there is one corresponding boundary curve or ‘operating characteristic curve’ (OCC; Embretson & Reise, 2000) described by the  $P_{ijx}^*$ . Figure 2.2 shows an example of the OCC for a five-category item under the GR model with  $a = 1.5$ ,  $b_1 = -1$ ,  $b_2 = 0$ ,  $b_3 = 1$ , and  $b_4 = 2$ . These curves are depicted according to the  $P_{ijx}^*$  values conditional on ability level ( $\theta_j$ ). The slope parameter ( $a_i$ ) determines the steepness of operating curves. Generally speaking, the higher the slope parameter, the steeper the curves. It should be noted that

under GR model, only one slope parameter is estimated for each item. Reflected by the OCC, all the curves converge but do not cross.

The threshold parameters ( $b_{ix}$ ) dictate the location of OCC. From Figure 2.2, one can see that the threshold represents a point on an ability scale where an examinee has a .50 probability of responding in or above a particular category. For instance, the first threshold in this example item  $b_1$  is -1, which means an examinee at ability level of -1 has .50 chance of obtaining a score of 1 or higher on this item. Moreover, the threshold parameters within an item are constrained to be ascending, that is,  $b_{i(x-1)} < b_{ix} < b_{i(x+1)}$ .

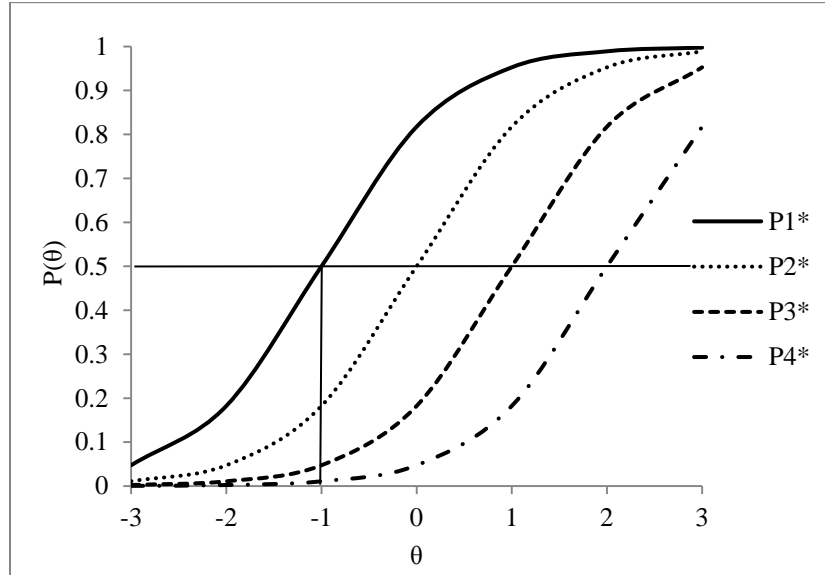


Figure 2.2 Operating characteristic curves for a five-category item under the GR model

Once these cumulative category response functions ( $P_{ijx}^*$ ) are estimated, the probability of obtaining a particular category is obtained by taking the difference between cumulative probabilities of adjacent categories. Thus, in GR model, the probability that an examinee responds to a particular category  $x$  ( $x = 0, 1, \dots, m_i$ ) on item  $i$  is given by

$$P_{ijx} = P_{ijx}^* - P_{ij(x+1)}^* \quad (2.3)$$

By definition, the probability of responding above the lowest category is 1.0, i.e.  $P_{i0}^* = 1$ , and the probability of scoring above the highest category is 0, i.e.,  $P_{ij(m_i+1)}^* = 0$ . Consider an item with five categories, four cumulative probabilities will be computed using Equation (2.2), that is,  $P_{ij1}^*$ ,  $P_{ij2}^*$ ,  $P_{ij3}^*$ , and  $P_{ij4}^*$ . Based on Equation (2.3), the probability of responding to a particular category ( $P_{ijx}$ ) can be calculated as follows:

$$\begin{cases} P_{ij0} = 1 - P_{ij1}^* \\ P_{ij1} = P_{ij1}^* - P_{ij2}^* \\ P_{ij2} = P_{ij2}^* - P_{ij3}^* \\ P_{ij3} = P_{ij3}^* - P_{ij4}^* \\ P_{ij4} = P_{ij4}^* - 0 \end{cases} \quad (2.4)$$

A graphical presentation of the probability of obtaining a particular score is called ‘category response curve’ (CRC, Embretson & Reise 2000). Figure 2.3 shows the CRCs for the same five-category item depicted in Figure 2.2. These curves are plotted based on each  $P_{ijx}$  conditional on ability. Similar to the OCCs, the shape and locations of the CRCs are also determined by the corresponding item parameters. Under GR model, the slope parameter determines the shape of the CRCs for middle categories. The higher the slope parameter, the narrower and more peaked the CRCs, indicating the response categories discriminate examinees with different ability levels fairly well. The threshold parameters determine the locations where the CRCs for middle response categories peak. Specifically, these category response curves peak in the middle of two adjacent threshold parameters. For example, as shown by the vertical line in Figure 2.3, the middle value for the first two thresholds ( $b_1 = -1$ ,  $b_2 = 0$ ) is -0.5, which is exactly the location where the response category curve for score category 1 peaks.

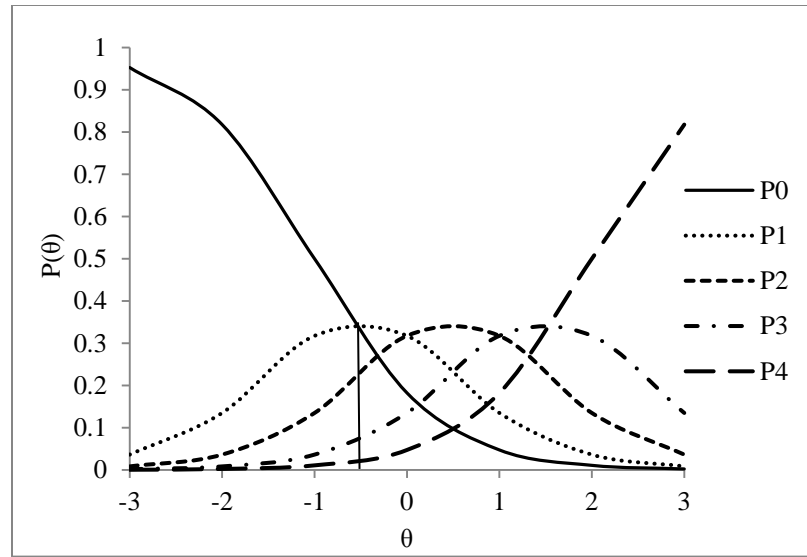


Figure 2.3 Category response curves for a five-category item under the GR model

## 2.2 MULTIDIMENSIONAL ITEM RESPONSE THEORY

### 2.2.1 Multidimensionality

One of central assumptions for conventional IRT models is unidimensionality which requires all the test items measure only one underlying latent trait. From the perspective of factor analysis, this assumption indicates that all the items have only one common factor. However, with the growing complexity in the process of test development, there are many testing situations that require the measurement of two or more latent traits. For example, ACT and SAT have multiple sections designed to measure different content domains or ability dimensions. When unidimensional IRT models are applied to such tests, distinctions between ability dimensions are ignored. This is particularly of great concern when underlying dimensions are not strongly correlated (Adams, Wilson, & Wang, 1997). Prior studies examining the consequences of fitting

unidimensional models to multidimensional data have demonstrated that unidimensional calibration usually provided biased parameter estimates (e.g., Ackerman, 1989; Ansley & Forsyth, 1985; Folk & Green, 1989; Way, Ansley, & Forsyth, 1988). In addition, a unidimensional approach fails to provide information on examinees' performance in each ability dimension. This information, however, is usually very useful in diagnosing examinees' problems in a particular area. In the context of school education, for instance, the information on various domains can help teachers, parents, and students to identify which area students have mastered or have problems with so that adjustments or remedies could be made accordingly.

In contrast to unidimensional IRT in which a single latent construct is assumed, MIRT accommodates multidimensional latent traits. MIRT can be considered as an extension of unidimensional IRT. Hence, the various forms of unidimensional models, either dichotomous or polytomous, can be generalized to a multidimensional context. Unlike unidimensional models where a single  $\theta$  is used, a vector of  $\boldsymbol{\theta}$  is identified in multidimensional models. For example, Reckase's (1997b) multidimensional extension of the unidimensional 3PL can be expressed as

$$p(x_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{(-\mathbf{a}_i' \boldsymbol{\theta}_j + b_i)}} \quad (2.5)$$

where

$x_{ij}$  is the response of examinee  $j$  responding to item  $i$ ;

$\boldsymbol{\theta}_j$  is an examinee's vector of abilities;

$\mathbf{a}_i$  is the vector of item discriminations;

$b_i$  is the location (difficulty) of item  $i$ ;

$c_i$  is the guessing of item  $i$ .

The notations of all the parameters are quite similar to those in the unidimensional 3PL model except that the ability and discrimination parameters are expressed as vectors.

Nevertheless, both conceptual and technical complexities have increased for MIRT (Reckase, 2009, p.63). One illustration of this complexity is the structure of multidimensionality at the item level. According to the pattern of the relations between items and ability dimensions, two types of multidimensionality have been identified (Adam, Wilson, & Wang, 1997; Hartig & Höhler, 2009). The first type is between-item multidimensionality or simple structure. Between-item multidimensionality is where multiple abilities are measured at the same time but each item only measures one ability. Between-item multidimensional models are also known as multi-unidimensional models. The second type of multidimensionality is within-item multidimensionality or complex structure. With this type of multidimensionality, some items measure more than one ability. Figure 2.4 illustrates a simple example of the two types of multidimensionality.

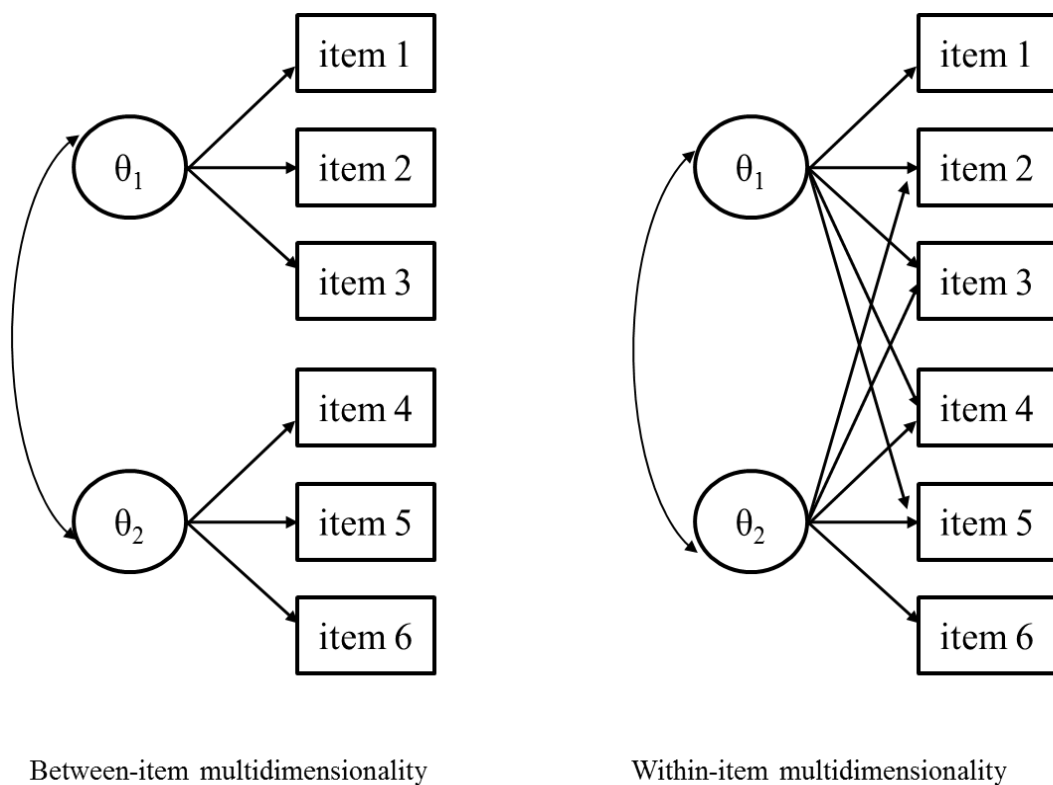


Figure 2.4 Two types of multidimensionality with two dimensions and six items

One advantage of between-item multidimensional models is that they are less complex than within-item multidimensional models both conceptually and statistically. In the between-item multidimensional models, the latent dimensions represent the combinations of all the abilities required to solve the respective items, regardless of how these abilities needed to be integrated to solve or respond to an item, which however would be required in the within-item multidimensional models. Therefore, they are more suitable in the context of gaining descriptive measures of performance in certain content areas (Hartig & Höhler, 2009). Between-item multidimensionality is probably more common in large-scale educational assessments where a test consists of several subtests with each focusing on a single ability and disjunctive clusters of items are designed to measure each ability (Sheng & Wikle, 2007). Furthermore, between-item multidimensionality is a common structure used in many other areas of social sciences such as health behavior. One example can be found in Allen and Wilson's (2006) study, where data from the Treatment Self-Regulation Questionnaire that measures self-determination was analyzed based on six subscales with each measured by a distinctive set of items. Consequently, MIRT model with between-item multidimensionality is considered for the present study.

### **2.2.2 Applications of MIRT**

There are various applications of MIRT in the testing field. As Reckase (2009) said, 'The application of MIRT to practical testing problems is a very active area of research' (p.75). One typical example of such applications is differential item functioning (DIF) analysis. DIF occurs when items behave differently for examinees from different subgroups of the population (e.g., gender) who have an equal ability level. A cause of DIF is the presence of multidimensionality in items. That is, the items measure some nuisance dimensions in addition to the goal dimensions



(Ackerman, 1992; Roussos & Stout, 1996). If subgroups of the population differ on these nuisance dimensions other than the goal dimensions, then the reported results may include bias. MIRT analysis explicitly allows items to differentiate examinees on multiple dimensions, which provides a natural means of identifying DIF or clarifying the reasons for a large DIF.

Another example of MIRT application is computerized adaptive testing (CAT). CAT is a measurement instrument that allows each examinee to receive a set of test items that are tailored to his/her own ability. IRT has provided possible algorithms to adaptively select and administer items to an individual examinee. When MIRT is accommodated within the context of CAT, an ability estimate in one dimension will provide clues about the examinee's standing along other dimensions due to the correlation among dimensions (Segall, 1996; 2000). This unique feature of MIRT may enhance the efficiency of an adaptive item selection as well as examinees' ability estimation compared to unidimensional CAT. Li and Schafer (2005) showed that MIRT CAT increased the accuracy of ability estimates, especially for low or high abilities, compared to a separate unidimensional CAT for each subscale. In addition, researchers have developed multidimensional linking procedures to map separate calibrations of MIRT tests to a common metric (e.g. Davey, Oshima, & Lee, 1996; Yao & Boughton, 2009). This makes it possible to develop large pools of calibrated items that can be used for CAT and for the construction of test forms that are parallel multidimensionally (Reckase, 2009).

Besides, the primary use of MIRT is to simultaneously and more precisely estimate multiple abilities for individuals. In the context of educational assessments, scores on each dimension can provide feedback to teachers and students in a classroom setting. This is also one of typical applications of between-item MIRT. With a conventional IRT approach, subscale scores can be obtained by applying a unidimensional IRT model to response data within each

subscale. The obtained subscale scores, however, are usually unreliable due to a small number of items included in each subscale. An ideal way to resolve the problem is to increase a test length. However, some practical considerations such as examinees' fatigue caused by a longer testing time might make it impossible (Wang, Chen, & Cheng, 2004). Hence, MIRT, which simultaneously estimate multiple abilities, provides a more realistic way to estimate subscale scores. The advantages of MIRT for estimating a subscale performance over the unidimensional IRT scoring for each dimension are: 1) MIRT considers the correlations across subscales or test batteries and hence, provides more accurate estimation on a subscale; 2) the correlations between latent dimensions can be directly modeled and estimated in MIRT; 3) all subscale scores obtained from MIRT are comparable across test forms and samples of examinees (Stone, Ye, Zhu, & Lane, 2010; Yao & Boughton, 2007).

Many studies have been conducted to examine the performance of MIRT in subscale ability estimation in comparison with the separate unidimensional calibration of each subscale. These studies have indicated that MIRT provides more accurate estimates since the correlations among subscales are taken into account (e.g., de la Torre, 2008, 2009; de la Torre & Patz, 2005; Hong, et al., 2010; Sheng & Wikle, 2007; Wang, et al., 2004; Yao & Boughton, 2007). In other words, item responses to other ability dimensions provide more information to estimate the ability for the current dimension. Meanwhile, most of these studies used a simulation approach to explore the performance of MIRT under various conditions. These conditions involved the number of dimensions, number of items in each dimension, correlation between ability dimensions, number of examinees, and number of item response categories. Although different forms of MIRT were employed in these simulation studies, most concluded that MIRT was most

efficient in a combination of following conditions: highly correlated abilities, a large number of dimensions, and a small number of items in each dimension.

### 2.2.3 Hierarchical MIRT Models

Even though subscale scores can provide diagnostic information on examinees' performance, it is still required in many circumstances to describe their competence on an aggregated level. For example, for an English proficiency test with three subtests: listening, reading, and writing, one may be interested in the overall scores to indicate a general proficiency in English in addition to scores on each domain. Traditionally, unidimensional IRT is implemented to estimate an overall ability. However, the estimates overlook the relationships among latent abilities and may not be valid because of the extent to which the unidimensional assumption is violated (de la Torre, 2009). Therefore, hierarchical structure models that incorporate both overall and specific domain abilities/subscales have been proposed. Higher-order IRT (HO-IRT) that involves broader and more general dimensions at a higher level and specific domain abilities at a lower order is one of typical hierarchical modeling of abilities. In HO-IRT, the first-order model is similar to between-item MIRT, i.e., each item is loaded on only one dimension. The second order is a linear model describing the relationship between an overall and multiple domain abilities, and these linear relationships determine the correlations among domain abilities.

As presented in Figure 2.5, the domain ability is expressed as a linear function of the overall ability, that is,  $\theta_{j(d)} = \gamma_{(d)}\theta_j + \varepsilon_{j(d)}$ , where  $\gamma_{(d)}$  ( $-1 \leq \gamma_{(d)} \leq 1$ ) is the latent coefficient in the regression of the ability on domain  $d$  predicted by the overall ability  $\theta_j$  which has a standard normal distribution (i.e.,  $\theta_j \sim N(0,1)$ );  $\varepsilon_{j(d)} \sim N(0, 1 - \gamma_{(d)}^2)$  is the error term that is

normally distributed with mean of zero and variance of  $1 - \gamma_{(d)}^2$ . Conditional on the regression coefficient and overall ability, the domain ability follows  $\theta_{j(d)} | \gamma_{(d)}, \theta_j \sim N(\gamma_{(d)}\theta_j, 1 - \gamma_{(d)}^2)$ . Through such a formulation, the marginal distribution of each domain ability is also standard normal (i.e.,  $\theta_{j(d)} \sim N(0,1)$ ). The correlation between the overall and domain abilities is indicated by  $\gamma_{(d)}$ , and correlation between the domain abilities, e.g.,  $\theta_{j(l)}$  and  $\theta_{j(k)}$ , is  $\gamma_{(l)} \times \gamma_{(k)}$ . Although  $\gamma_{(d)}$  can be mathematically negative, it is usually expected to be nonnegative because domain abilities are typically positively correlated to the overall ability (de la Torre & Song, 2009).

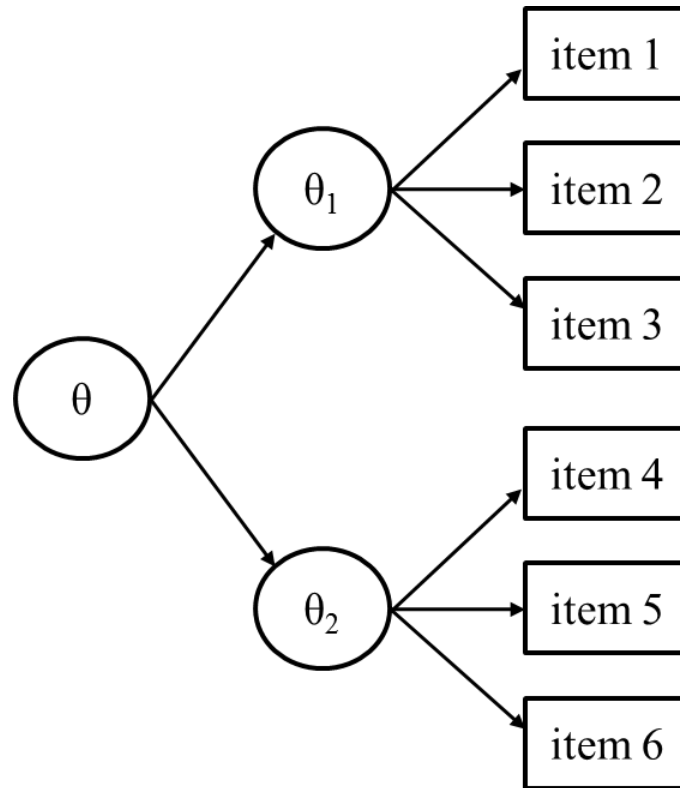


Figure 2.5 Illustration of a HO-IRT model

Using a multidimensional 3PL model at the lower order, de la Torre and Song (2009) conducted a simulation study to evaluate the performance of the HO-IRT model in ability estimation as a function of number of domain abilities, number of items in each domain,

correlation between domains, and number of examinees. The HO-IRT estimates of both overall and domain abilities were evaluated and compared to the unidimensional IRT calibrations in terms of the correlation between the true and estimated abilities, the posterior variance, and the mean squared error (MSE). For domain abilities, the estimates produced by the HO-IRT model were found to be less biased and more efficient than those derived from the unidimensional calibration of each domain. The optimal condition of the HO-IRT model was multiple shorter tests measuring multiple highly correlated dimensions. These results are quite consistent with studies using MIRT models (e.g., de la Torre, 2008, 2009; de la Torre & Patz, 2005; Hong, et al., 2010; Sheng & Wikle, 2007; Wang, et al., 2004; Yao & Boughton, 2007). With respect to overall ability estimates, the study found that the HO-IRT model showed less bias, and was generally more efficient than the unidimensional IRT model, particularly when the domain abilities were uncorrelated.

Based on de la Torre and Song's (2009) study, de la Torre and Hong (2010) further investigated the performance of HO-IRT models in item parameter recovery in addition to the ability estimation, particularly for smaller sample sizes. The accuracy of item parameter estimates was evaluated from different perspectives. At parameter level, root mean square error (RMSE) of the item parameter estimates was computed. At item level and test level, root mean square difference (RMSD) between the true and estimated item characteristic curves and RMSD between the true and estimated test characteristic curves were computed respectively. All these statistics provided evidence that the HO-IRT model improved the item parameter estimation compared to the separate calibration of each dimension. The pattern of improvement was best observed at test level. Findings from RMSD of test characteristic curves indicated the HO-IRT model was more advantageous over the unidimensional IRT approach in estimating item

parameters when the sample size was relatively small, the test was relatively shorter, and the number of domains was higher.

According to different assumptions on the structural relationships between an overall and domain abilities, there are some other forms of hierarchical models have been proposed and examined. For example, in addition to HO-IRT, Sheng and Wikle (2008) proposed a MIRT model with a different hierarchical structure. Unlike HO-IRT which assumes that each domain ability is a linear function of a general ability, this alternative hierarchical model, like the simple averaging method, assumes a general ability is a linear combination of specific abilities. Furthermore, unlike HO-IRT in which the correlations between domain abilities are manifested by the overall ability at the second order, this alternative hierarchical structure directly models the correlations between domain abilities at the first order (i.e., multi-unidimensional model at the first order). This alternative hierarchical model was compared to HO-IRT along with several other models on item parameter recovery and model fit. The simulation studies indicated the alternative hierarchical model performed better than the HO-IRT model in recovering item parameters under some simulation conditions. However, both hierarchical models provided a better model fit compared to the unidimensional and multi-unidimensional IRT models.

The recent study conducted by Yao (2010) applied the maximum information function to obtain an overall ability from domain abilities. This approach does not assume a linear relationship between the overall and domain abilities. In this method, MIRT is applied first for the domain ability estimation, and then the overall ability is obtained by the maximum information function. To avoid the situation where the fitting model is favored due to its identicalness to the model for data generation, the author simulated item responses based on both HO-IRT and MIRT models. The two models were compared on the recovery of item parameters,

overall ability, and domain abilities under a variety of conditions. The results showed that both models performed similarly on recovering the item parameters and domain abilities. The performance on the overall ability recovery was dependent on the simulating model and magnitude of the correlations among domains. When the fitting model conformed to the true simulating model, both methods performed similarly well on the overall score recovery. When the fitting model was not the same as the simulating model, the maximum information method seemed to perform better than the HO-IRT model, particularly when the correlations among domain abilities were low. However, in terms of classification of examinees based on the overall ability, the MIRT with maximum information estimation gave the worst match at the two ends when the HO-IRT model with low correlations was the true model. In general, the two hierarchical models performed similarly as long as the model was correctly specified.

Although the MIRT models specified by previous studies may vary to each other in terms of model structure or model types (e.g., 2PL vs. 3PL), most have indicated that MIRT models outperformed unidimensional IRT models in the estimation of both item and person parameters.

## **2.3 COMPARISONS BETWEEN IRT AND CFA**

### **2.3.1 Brief Introduction to CFA**

CFA is one of core techniques of structural equation modeling (SEM). A standard CFA model is specified to reflect a hypothesized relationship among observed variables (indicators), factors that represent constructs intended to be measured by the indicators, and measurement errors that denote all the unique sources of variance not explained by the factors. That is to say, a researcher

must have a prior postulation on the number of factors and relationship patterns between factors and indicators based on past evidence and theory. This hypothesis-driven feature makes CFA different from its counterpart, exploratory factor analysis (EFA), which is primarily used as a descriptive or exploratory technique to determine the number of factors and detect which indicators are reasonable measures of the underlying construct. As one of the most commonly used statistical techniques in applied research, CFA has various applications such as psychometric evaluation of test instruments, construct validation, method effects identification, and measurement invariance evaluation (Brown, 2006).

Using a variance-covariance matrix of indicators as input, CFA aims to reproduce this sample matrix by estimating the parameters of a measurement model. The fundamental parameters in all CFA models include factor loadings, factor variance and unique variance. Factor loadings are used to quantify the effects of underlying factors on observed variables. They can be generally interpreted as regression coefficients. A factor variance represents the variability or dispersion of a factor, and a unique variance typically refers to a measurement error. A correlation matrix, which is a standardized form of a variance-covariance matrix, can be also used as input. This will result in the standardized estimates for factor loading parameters, i.e., standardized regression coefficients.

In order to estimate the parameters in CFA, a model must be identified (Jöreskog, 1969). There are two necessary requirements for any CFA models to be identified. First, every latent variable scale must be identified. Latent variables are unobservable by nature, and thus have no defined metric. In order to set a unit for a latent variable, one can either fix one of the factor loadings at 1 or fix the variance of the factor to a constant (usually 1.0). Either scaling factor solution has its consequences. If one of the factor loadings is fixed at 1, the metric of the factor



will be the same as one of its indicators, i.e., unstandardized factor loading on one indicator. Alternatively, fixing a factor variance at 1 will result in the factor being measured on a standardized normal scale, i.e.,  $N \sim (0, 1)$ . Nonetheless, both scaling solutions generally result in the same overall model fit of the same data (Kline, 2005). Second, the number of freely estimated parameters (e.g., factor loadings, factor variances/covariances) must not exceed the number of unique elements of a variance-covariance matrix. The number of unique elements in a variance-covariance matrix equals  $p(p+1)/2$ , where  $p$  is the number of observed variables/indicators. For example, for a CFA model with two correlated factors and each factor with four indicators, the second condition is met because the number of observations ( $8(8+1)/2=36$ ) in the input matrix is less than the number of estimated parameters (i.e., 8 factor loadings+8 error variances+1 factor covariance=17 if the variance of each factor is constrained to 1). Therefore, for a CFA model with a single factor, a minimum of three indicators is required in order to have the model just identified.

Like HO-IRT, CFA can incorporate higher-order factors to account for the correlations among lower-order factors. This hierarchical structure is originally popular in intelligence research where a broader dimension of general ability ( $g$ ) influences more specialized facets of ability (e.g., verbal comprehension, perceptual organization, memory, etc.). This broader dimension is known as a second-order factor. Although a higher-order factor analysis can proceed to third order or beyond, such analysis is seldom used in applied literature (Brown, 2006). Furthermore, more than one factor can be specified at the higher order depending on the theory about the construct structure. The rules of identification used for a standard first-order CFA model can be generalized to higher-order solutions. For example, in order for a CFA model with a single second-order factor to be identified, there must be at least three first-order factors.

HO-IRT and second-order CFA are, in fact, equivalent in the model structure although they were developed from different frameworks. An alternative hierarchical model that is mathematically related to higher-order IRT or CFA models is bi-factor model. In a bi-factor model, there is a single general factor for all items plus one or more orthogonal specific factors for some or all of items. Figure 2.6 shows an example of bi-factor model with a general factor ( $\theta$ ) and two specific factors ( $\theta_1$  and  $\theta_2$ ). As one can see from Figure 2.6, the general factor has a direct effect on all the items but not on the specific factors. In contrast, in a higher-order factor model, the general factor has a direct effect on the specific factors but an indirect effect on all items through the specific factors. Nevertheless, the two types of models are related and can be transformed into each other with some constraints (Yung, Thissen, & McLeod, 1999). The choice between the higher-order and bi-factor models largely relies on the theoretical hypothesis of the relationship between latent dimensions and respondents' performance. The higher-order model assumes that the variability in respondents' performance is accounted for solely and directly by specific factors while the bi-factor model assumes that the variability is accounted for directly by both general and specific factors. Moreover, the higher-order model seems more flexible than the bi-factor model since it can be formulated to have more than one general factor, and to allow nonlinear relationships between a general factor and specific group factors (de la Torre & Song, 2009).

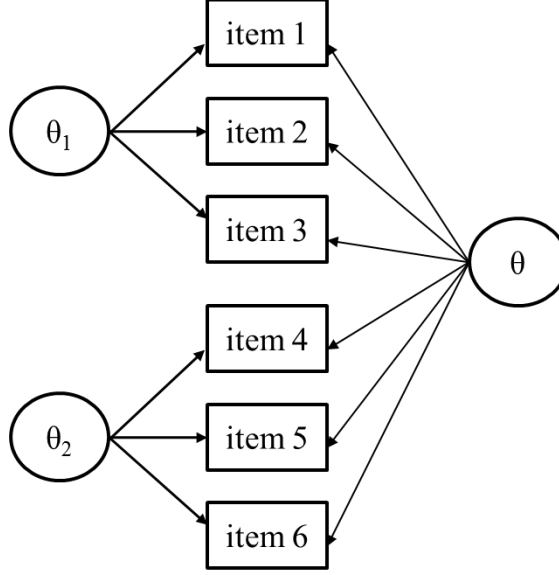


Figure 2.6 Example of a bi-factor model

### 2.3.2 Comparisons between IRT and CFA

CFA was initially developed for continuous variables, but has been extended to include categorical variables, which makes it possible to offer a comparable analytical framework to IRT. For a categorical variable, CFA assumes that the observed categorical response ( $x_{ij}$  for individual  $j$  on item  $i$ ) is a manifestation of an underlying continuous response. For example, for an item, ‘I’m satisfied with my life in general’, a dichotomous response ‘yes’ or ‘no’, in fact, represents an underlying latent construct that has a continuous distribution. These underlying continuous variables ( $y_j$ ) were assumed to be normally distributed and usually follow

$$y_j \sim N(0, \Sigma) \text{ for each individual } j \quad (2.6)$$

where  $\Sigma$  denotes the variance and covariance matrix of the underlying continuous variables. In practice, a tetrachoric correlation is typically used for dichotomous responses (a polychoric correlation is used for polytomous responses), which is an estimate of correlation of the normally

distributed continuous latent variables based on the observed dichotomous responses. Moreover, the covariance matrix  $\Sigma$  is defined as

$$\Sigma = \Lambda\Phi\Lambda' + \Psi \quad (2.7)$$

where  $\Lambda$  is a matrix of factor loadings,  $\Phi$  is a matrix of variances and covariances among the factors, and  $\Psi$  is a diagonal matrix containing unique variances.

Each continuous variable  $y_{ij}$  under the CFA models are dichotomized by the following criteria

$$x_{ij} = 1 \text{ if } y_{ij} \geq \tau_i \text{ or } 0 \text{ otherwise.} \quad (2.8)$$

where  $\tau_i$  is called the threshold parameter for item  $i$ , which corresponds to the point on a continuous latent scale that separates the two response categories. When there is more than one item, however, this categorization procedure depends not only on the marginal latent response distributions but also on the correlations among these latent distributions. Therefore, as the number of items increases, the estimation process becomes more complex.

When all the items are assumed to measure one common factor, it follows

$$y_{ij} = \lambda_i \eta_j + \varepsilon_{ij} \text{ with } \varepsilon_{ij} \sim N(0, \sigma_i^2) \quad (2.9)$$

where  $\sigma_i^2 = 1 - \lambda_i^2$ ;  $\lambda_i$  is the estimated factor loading of item  $i$ ;  $\eta_j$  is called factor score, which corresponds to the latent trait ( $\theta_j$ ) in IRT; and  $\varepsilon_{ij}$  is the estimated unique variance. As seen in Equation (2.9), each latent continuous variable can be expressed as a function of a latent factor and some unique errors. Based on Equations (2.8) and (2.9), model-implied estimates are derived and compared to observed data to evaluate a goodness of model fit. Although different parameters seem to be used for IRT and CFA, a formal relationship of the two classes of models has been established. Takane and de Leeuw (1987) demonstrated the equivalence of marginal probabilities of the two-parameter normal-ogive IRT and CFA models for dichotomous data. From Equation (2.9), they proved that it can be derived that

$$p(x_{ij} = 1|\theta_j) = \Phi\left(\frac{\lambda_i\theta_j - \tau_i}{\sigma_i}\right) \quad (2.10)$$

where  $\Phi(\cdot)$  is a normal-cdf function.

If setting

$$a_i = \frac{\lambda_i}{\sigma_i} \quad (2.11)$$

$$b_i = \frac{\tau_i}{\lambda_i} \quad (2.12)$$

Then Equation (2.10) can be rewritten as

$$p(x_{ij} = 1|\theta_j) = \Phi[a_i(\theta_j - b_i)] \quad (2.13)$$

Equation (2.13) is exactly the form of two-parameter normal-cdf model in which  $a_i$  and  $b_i$  are denoted as discrimination and difficulty parameters respectively. Note  $\sigma_i$  in Equation (2.11) seems to be an additional parameter in CFA, but it is actually a function of  $\lambda_i$ , i.e.,  $\sigma_i = \sqrt{1 - \lambda_i^2}$ . Therefore, the factor loading and threshold parameters in CFA can be transformed to the discrimination and difficulty parameters in IRT, and vice versa. When item responses are polytomous, Equation (2.12) can be easily generalized to multiple threshold parameters. That is, multiple thresholds in CFA can be transformed to the corresponding threshold parameters in graded response model as

$$b_{ix} = \frac{\tau_{ix}}{\lambda_i} \quad (x = 1, \dots, m_i \text{ number of score levels}) \quad (2.14)$$

Furthermore, when multiple factors are identified in CFA, which is equivalent to multidimensional IRT, similar transformations as those in unidimensional IRT can be performed. These transformations have been identified and used in many previous research (e.g., Finch, 2010; Glöckner-Rist & Hoijsink, 2003; Knol & Berger, 1991; McLeod & Swygert, 2001).

Despite the mathematical equivalence between CFA and IRT, there are some differences between the two techniques. Most of these differences are ultimately related to applications. First, the focus on the characteristics of the input variables is different between the two methodologies (Reckase, 1997a). CFA uses a correlation or covariance matrix of observed variables but ignore some individual characteristics of each variable. IRT, on the contrary, regards characteristics of individual items such as difficulty and discrimination as an important component in an analysis. Particularly, the guessing parameter in dichotomous IRT is usually very hard to be incorporated in CFA. Second, two methodologies have different perspectives for assessing a goodness of model fit. CFA considers a model fit as a global measure of the empirical relationships among all observed variables compared to the relationships implied by the structure of a theoretical model (model-implied covariance matrix). IRT, on the other hand, tends to establish a model fit at an item level as well as a person level. Therefore, a model fit in IRT is concerned with discrepancies on a single ill-modeled item or for a particular range of abilities (Reckase, 1997a). From a practical sense, IRT is more appropriate if a researcher is interested in individual item characteristics or scoring for each subject. CFA may be the choice if research questions focus on the structural relationship between constructs and responses (Wirth & Edwards, 2007).

Another practical distinction between CFA and IRT is how the estimation is routinely conducted (Edwards, 2010). Takane and de Leeuw (1987) explicitly discussed the difference between the two methodologies on the marginalization being performed. In the IRT formulation, the dichotomization step is first performed conditionally on ability and then the marginalization step is undertaken. In CFA, the marginalization step is performed on continuous underlying variables, followed by the dichotomization step. This marginalization issue has also posed a

challenge for both approaches in parameter estimation. Wirth and Edwards (2007) pointed out that a numerical integration over factors is usually required for IRT while integration over items is needed for CFA. The computational difficulty in an analytical integration limits the number of factors or items that can be used for both approaches. The difference in marginalization to some extent determines the choice of estimation methods under the two models.

## **2.4 ESTIMATION METHODS IN IRT AND CFA**

### **2.4.1 Estimation Methods in IRT**

#### **2.4.1.1 Marginal Maximum Likelihood (MML) vs. Unweighted Least Square (ULS)**

Under the IRT framework, one of the most commonly used estimation methods is MML with an expectation-maximization (EM) algorithm (MML/EM; Bock & Aitkin, 1981). Unlike most typical estimation methods in CFA where the correlations or covariances among items are used as an input for analysis, the MML/EM method uses all response vectors and therefore is referred to as a full-information estimation. This approach attempts to integrate over person-specific parameters and estimate item parameters in the marginal distribution. That is, it tends to find the item parameter estimates that maximize the likelihood in which person parameters have been removed (Wirth & Edwards, 2007). With an EM algorithm, the expected number of persons at each ability level and the expected number of persons passing each particular item are first computed in the expectation stage, and then in the maximization step, item parameter estimates are found to maximize likelihoods using those expected values. This iterative process continues until some convergence criterion is met. This estimation method is readily applicable to all types

of IRT models, including multidimensional models (Embretson & Reise, 2000). For instance, Bock, Gibbons, and Muraki (1988) used this estimation method for an exploratory MIRT model. The MML/EM algorithm has been implemented in many popular IRT computer programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), both of which can be used for unidimensional IRT, and TESTFACT (Bock, et al., 2002) which can be used for MIRT in addition to unidimensional IRT. Nevertheless, the MML/EM approach has struggled with the integration over the number of latent abilities, i.e., the number of latent traits that can be estimated is limited. The issue becomes particularly salient for research areas that measure multiple factors in one inventory such as psychological tests (Wirth & Edwards, 2007).

Another estimation method for both unidimensional and multidimensional IRT is NOHARM, which was developed by McDonald (1997) and programmed by Fraser and McDonald (1988). Distinctive from the full-information factor analytic approach, NOHARM relies on pairwise information given by the proportions of examinees passing any pair of items. Although the estimation method in NOHARM is based on the normal-ogive MIRT, it actually fits data using a polynomial approximation to the normal ogive. McDonald (1997) pointed out that a polynomial term up to a cubic was sufficient for estimation in most cases. The actual estimation of parameters is performed using ULS based on the criterion of minimizing the differences between observed and model-predicted proportions of examinees correctly responding to two items across all possible pairs of items.

As both TESTFACT and NOHARM can be used to analyze multidimensional response data, the two programs have been compared in many simulation studies. It is also necessary to recognize that, although both programs are developed under the IRT framework, the estimation methods are in fact heavily borrowed from factor analysis and SEM approaches (Reckase,



1997a). Therefore, both procedures have also been examined in assessing dimensionality and a factor structure in addition to item parameter estimation. The findings of most previous studies indicated that both programs performed similarly in item parameter recovery and in identifying factor structures (e.g., Béguin & Glas, 2001; Stone & Yeh, 2006; Tate, 2003). However, Gosz and Walker (2002) found TESTFACT performed better when the correlation between items was large, and when a test contained a greater number of multidimensional items while NOHARM performed better in the opposite conditions. It is difficult though to generalize the results from these studies to the analysis of new datasets since different conditions, such as the number of examinees and the number of items, are involved in different studies. In addition, there are some practical limitations for either or both programs. First, neither TESTFACT nor NOHARM provides an estimation of guessing parameters, but they have the option of entering fixed guessing parameters into the estimation procedure. These values can be obtained using other software such as MULTILOG, and then they can be included in the estimation of a factor model. Secondly, NOHARM does not provide ability estimates ( $\theta$ -vectors), but TESTFACT does. Thirdly, although both procedures can be used for assessing a scale's dimensionality and factor structure in both exploratory and confirmatory ways, TESTFACT can only perform CFA based on a bi-factor model. Finally, both programs are confined to analyzing dichotomous responses when the data are multidimensional.

#### **2.4.1.2 Markov chain Monte Carlo (MCMC) Simulation**

Another estimation method that has become increasingly popular in IRT is MCMC. MCMC is a method of simulating random samples from any theoretical or target distribution so that the features such as mean and variance of a target distribution can be estimated (Patz & Junker, 1999a). As a sampling-based estimation method, MCMC avoids the analytic integration that is

often used for a standard estimation of IRT parameters such as MML, and therefore can deal with more complex models.

MCMC is an estimation strategy that is rooted in a perspective of Bayesian inference. Bayesian statistics is based on the idea of expressing the uncertainty of unknown parameters in terms of probabilities. In the Bayesian framework, an unknown population parameter is assumed to be a random variable that has a certain distribution. We formulate a prior distribution based on our beliefs or prior knowledge on the unknown parameter distribution before observing data. This prior distribution will be updated by the observed data through the likelihood to form a posterior distribution for the parameter. A posterior distribution, which takes into account both a prior distribution and data, reflects a probability of the unknown parameter. Mathematically, Bayes theorem is applied to obtain a posterior distribution. Let  $\theta$  denotes a vector of unknown parameters, and  $D$  denotes observed data, then

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)} = \frac{p(\theta)p(D|\theta)}{p(D)} \quad (2.15)$$

where  $p(\theta, D)$  is the joint probability distribution for  $\theta$  and  $D$ ;  $p(\theta)$  is the prior distribution of parameters and it represents researchers' prior information or belief about  $\theta$ ;  $p(D|\theta)$  is the likelihood function of the data given the parameters;  $p(D) = \int p(\theta)p(D|\theta)d\theta$  is the marginal or unconditional probability of data across all possible values of  $\theta$ .

$p(D)$  is usually treated as a constant for fixed data because it is only a function of data and does not depend on  $\theta$ . This constant is also referred to as 'normalization constant' (Gilks, Richardson, & Spiegelhalter, 1996) as it scales the likelihood  $p(D|\theta)$  to have a proper probability density. However, in many situations, computing the normalization constant is not feasible or simply unnecessary (Rupp, Dey, & Zumbo, 2004). Therefore, an almost equivalent form of Equation (2.15) is obtained by omitting  $p(D)$  as follows,

$$p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta}) \quad (2.16)$$

This indicates the posterior distribution of the given data is proportional to the product of the prior distribution and likelihood of the data.

One of the main goals of Bayesian inference is to sample from a posterior distribution to estimate population parameters (e.g., means and variance), to construct credible intervals (i.e., Bayesian confidence intervals), and to obtain Bayesian posterior  $p$  values for hypothesis testing (Rupp, et al., 2004). When a posterior density function has a close form, sampling from the posterior is direct, and calculation of mean, variance, or credible intervals is straightforward. However, it is often the case, especially in complex models, that posterior distributions do not have analytical forms, and thus a direct sampling from them is impossible (as indicated by Equation 2.16). MCMC simulation becomes invaluable in this situation because it provides an iterative process of drawing samples from a distribution that is close enough to the posterior distribution. MCMC can be thought of as Monte Carlo integration using Markov chains (Gilks, et al., 1996). Monte Carlo integration works by drawing samples from a posterior distribution and then computing averages to approximate the expectations. Monte Carlo integration replaces what can be a very complex analytic integration such as MML with simple computations (Wirth & Edwards, 2007).

Constructing a Markov chain with its stationary distribution being the posterior distribution is the key to MCMC estimation. A Markov chain involves a sequence of random variables for which the state at any time  $t$ , say  $\theta^t$ , is sampled from a distribution  $p(\theta^t|\theta^{t-1})$  which depends only on its state at the time  $t-1$  and not any prior states of the chain. Subject to some general conditions, a chain will eventually converge to a stationary distribution, which does not depend on time or its initial values. When a Markov chain is stationary, sampling from

that chain will approximate sampling from the target posterior distribution. The samples from the chain then can be used to estimate the population parameters of interest.

There are different types of algorithms used to construct Markov chains within MCMC. The two commonly used with Bayesian IRT estimation are Gibbs sampler and Metropolis-Hastings (MH) algorithm. The Gibbs sampler is appropriate when the full conditional distribution of each parameter is known, and sampling from it is easy. A full conditional distribution represents the posterior distribution of a single parameter given all other parameters and data. Thus, sampling each parameter individually with respect to its conditional distribution corresponds to sampling from a joint posterior distribution. The Gibbs sequence involves sampling a value of one parameter conditional on other parameters and then drawing a value of subsequent parameters conditional on the newly drawn values of previous parameters. The process continues iteratively until convergence. For example, for  $p$  unknown parameters  $\theta_1 \dots \theta_p$ , the Gibbs sampler proceeds as follows:

1. Start with possible values  $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)}$  at iteration 0.
2. Given the data (D), generate new values for the parameters at iteration 1,

$$\theta_1^{(1)} \sim p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_p^{(0)}, D)$$

$$\theta_2^{(1)} \sim p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_p^{(0)}, D)$$

$\vdots$

$$\theta_p^{(1)} \sim p(\theta_p | \theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(1)}, \dots, \theta_{p-1}^{(1)}, D);$$

At any iteration  $t$ ,

$$\theta_1^{(t)} \sim p(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_p^{(t-1)}, D)$$

$$\theta_2^{(t)} \sim p(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_p^{(t-1)}, D)$$

$\vdots$

$$\theta_p^{(t)} \sim p(\theta_p | \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_{p-1}^{(t)}, D).$$

3. Repeat the procedure until the chain has converged.

When the conditional distributions are not of known functional forms, MH is used as an alternative to Gibbs sampling. In fact, Gibbs sampler has been widely viewed as a special case of MH algorithm (Chib & Greenberg, 1995; Gilks, et al., 1996). MH algorithm uses a proposal distribution from which a candidate value of a parameter ( $\theta'$ ) is sampled. The proposal distribution  $q(\theta' | \theta^{t-1})$  is pre-specified with the current state of the parameter  $\theta^t$  depending on the previous state  $\theta^{t-1}$ . For example, the proposal distribution might be a normal distribution with the mean determined by the previous state of the parameter in the chain and a specified variance. The candidate value sampled from this proposal distribution is accepted with the probability factor defined as

$$\alpha = \min\left(\frac{p(\theta')q(\theta^{t-1}|\theta')}{p(\theta^{t-1})q(\theta'|\theta^{t-1})}, 1\right) \quad (2.17)$$

If the candidate draw is accepted, it becomes the current state of the chain ( $\theta^t = \theta'$ ). If the draw is not accepted, the old value will be kept ( $\theta^t = \theta^{t-1}$ ). For  $p$  unknown parameters  $\theta_1 \dots \theta_p$ , the MH algorithm proceeds as follows:

1. Start with  $\boldsymbol{\theta} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})$
2. For iteration  $t$ , draw  $\boldsymbol{\theta} \sim q(\cdot | \boldsymbol{\theta}^{(t-1)})$
3. Compute the probability factor

$$\alpha = \min\left(\frac{p(\boldsymbol{\theta})q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta})}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})}, 1\right)$$

4. With probability  $\alpha$ , set  $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$ , otherwise set  $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ .

5. The procedure is repeated until convergence.

Gibbs sampler and MH algorithm are two basic methods used in MCMC. There are other more complex techniques being developed, many of which are modification, generalization, or hybrids of the two methods (Roberts, 1996).

When MCMC algorithm is used, it is very important to assess the convergence of Markov chain in that random draws are from the posterior distribution of interest only when the chain has converged. In common practice, there are some informal but quite useful approaches to assess convergence. Two typical examples are time-series plots and autocorrelation plots. A time-series plot, also called a history plot, provides the sampled values of a parameter at each iteration, indicating how much the draws for a parameter varies around the space at each iteration. A stable, well-mixed time-series plot is often an indication of convergence. On the other hand, when some trend in a sample space has been observed, it might be a sign of non-convergence. If multiple chains are used for estimation, then the history plots can provide a qualitative determination of whether the multiple chains, started at dispersed initial values, have merged and are drawing from a common distribution (Cowles, 2004). A convergence is often reflected by a large overlap in history plots.

An autocorrelation plot provides the correlation of the sequential draws of a parameter in the chain at varying time lags. The reason why examining autocorrelation is important is that high autocorrelations within chains can cause slow convergence. In other words, reducing autocorrelations can improve the efficiency of MCMC. Typically, an autocorrelation can be reduced by thinning a Markov chain, which is a procedure to take iterations at an even interval, for example, keep every 5<sup>th</sup> or 10<sup>th</sup> value to construct a posterior. By thinning the Markov chain, the dependency of samples can be reduced. In addition, re-parameterizing the model to reduce

the correlation between parameters may reduce an autocorrelation. Both time-series plots and autocorrelation plots can be easily constructed in WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2000), a program that implements MCMC estimation for Bayesian inference.

In addition, various formal diagnostic criteria have been proposed to evaluate the likelihood of convergence. Cowles and Carlin (1996) and Brooks and Roberts (1998) provided extensive review of convergence diagnostics that were developed in MCMC literature. One debate over convergence diagnostics involves whether to estimate one very long Markov chain or to run several relatively shorter chains. The common practice seems to estimate a few independent chains with different starting values (Edwards, 2010). If all the chains escape the influence from their starting points and converge to the same stationary distribution, then there is a strong likelihood of convergence. Gelman and Rubin's (1992) criterion is an example of convergence diagnostics using multiple chains. This diagnostic test is based on the comparison between a between-chain variance and a within-chain variance. Based on these quantities, the statistic is expressed as

$$\hat{R} = \frac{n-1}{n} + \frac{1}{n} \frac{B}{W} \quad (2.18)$$

where  $B/n$  is the variance between the means from the  $m$  independent chains with the length  $n$  (between-chain variance), and  $W$  is the average of the  $m$  within-chain variance.

During the initial sampling,  $B$  would be much larger than  $W$  because starting points are over-dispersed relative to a target distribution (Cowles & Carlin, 1996). Once the convergence is reached,  $B$  and  $W$  should coincide and  $\hat{R}$  should approximately equal one. When the value of  $\hat{R}$  of all parameters is close to 1, we may conclude that Markov chains converge; otherwise the algorithm may fail to converge. Gelman-Rubin statistic is reported by WinBUGS through a plot called Brooks-Gelman-Rubin (BGR) diagnostic plot which includes not only Gelman-Rubin

statistic, but also the average width of the 80% intervals within the individual estimations and the width of the 80% interval of the pooled estimations. Brook and Gelman (1998) emphasized that one should be concerned both with convergence of  $\hat{R}$  to 1, and with convergence of both pooled and within interval widths to stability. One drawback of Gelman and Rubin statistic is that its value depends greatly on the choice of initial values.

Various other techniques exist for assessing convergence in MCMC estimation, but there has been no single absolute measure that can be fully relied upon. Therefore, examination of different convergence criteria is recommended in common practice (Sinharay, 2004; Edwards, 2010). It is a researcher's responsibility to choose from and make use of multiple criteria to warrant a valid conclusion about a model. This procedure also adds complexity to MCMC estimation.

There are some other questions in the practice of applying MCMC. For example, at what point do we know that a stationary distribution has converged? In MCMC simulation, it usually takes a certain number of iterations for a Markov chain to reach convergence. Therefore, early iterations before a Markov chain has converged are usually discarded, which is referred to as a burn-in phase. Raftery and Lewis (1992) recommended the length of a burn-in should be at least as large as the distance between samples needed to have an autocorrelation of zero. A more conservative recommendation is to discard the first half number of iterations (Gelman, Carlin, Stern, & Rubin, 2004, p. 295). However, Geyer (1992) suggested setting a burn-in to between 1% and 2% of the total number of iterations as he argued that the actual burn-in is usually less than 1% of the total length of a sufficiently long estimation to obtain the adequate precision. In addition, the number of burn-in iterations also depends on situations where MCMC is applied.



For instance, in the context of IRT, it has been common to discard the first 500 or so iterations as a burn-in period (Kim & Bolt, 2007).

Another practical issue that needs to be considered is how many iterations are needed to summarize a posterior distribution after convergence. Since a posterior distribution is constructed from samples, the errors of estimates based on a posterior distribution can be attributed not only to the standard errors of a point estimate that are reflected by the standard deviation of a posterior, but also to sampling errors, referred to as Monte Carlo errors (Kim & Bolt, 2007). As a rule of thumb, a simulation should be performed until the Monte Carlo error for each parameter is less than about 5% of a sample standard deviation (Spiegelhalter, Thomas, Best, & Lunn, 2003).

Beginning with the work of Albert (1992) on using Gibbs sampling to estimate two-parameter normal-ogive model, followed by Patz and Junker (1999a, 1999b) who developed MH sampling algorithms to estimate two-parameter logistic models and generalized partial credit models, the Bayesian approach with MCMC estimation has been widely used in IRT. A large number of studies have compared MCMC estimation to MML/EM estimation, especially for unidimensional IRT models. Most of these studies have indicated that the two methods yielded similar estimates of item and person parameters (e.g., Baker, 1998; Béguin & Glas, 2001; Kieftenbeld & Natesan, 2012; Kim, 2001; Wollack, Bolt, Cohen, & Lee, 2002). Many researchers (e.g., Baker & Kim, 2004; Wollack et al., 2002) also suggested that MCMC approach is likely to be more useful for more complex models such as multidimensional polytomous models.

Edwards (2010) conducted a study in which MCMC estimation was compared to MML/EM using polytomous MIRT models. The model consisted of four correlated dimensions

with each item only measuring one dimension. Since there is no available software implementing MML/EM procedure for polytomous MIRT, the author performed MML/EM calibration separately for each dimension using MULTILOG. This is also a reason why studies comparing MCMC with other estimation methods such as TESTFACT or NOHARM in the MIRT context are restricted to dichotomous MIRT models (e.g. Béguin & Glas, 2001; Bolt & Lall, 2003), although most of them have indicated that the parameters estimated by MCMC do not importantly differ from the estimates from the other two programs. In Edwards's (2010) study, he found that MCMC was slightly superior to MULTILOG in the recovery of item parameters, and this advantage was found to be more pronounced for smaller samples. The more accurate estimates derived from MCMC should be attributed to the incorporation of information provided by multiple correlated dimensions. Due to its flexibility and easy generalization to complex high-dimensional models, the Bayesian approach with MCMC estimation has been commonly used for MIRT models (e.g., de la Torre & Patz, 2005; Yao & Schwarz, 2006) as well as other complex IRT models such as testlets models (e.g., Bradlow, Wainer, & Wang, 1999) and multilevel IRT models (e.g., Fox & Glas, 2001).

Admittedly, MCMC has its own disadvantages. One typical shortcoming is that the method is computationally intensive, and therefore can be very time-consuming, particularly when the number of parameters is large. Moreover, the available software specifically designed for MCMC implementation is sparse. WinBUGS (Spiegelhalter, et al., 2003) is one typical production program developed for MCMC. Although there are some other programs such as BMIRT (Yao, 2003) and MultiNorm (Edwards, 2005) developed within the IRT framework, they have not been widely recognized and used. A comparison of these programs might be one

of directions for future research. Nonetheless, MCMC is a very promising estimation method to deal with more complex models.

## 2.4.2 Standard Estimation Methods in CFA

Under the CFA framework, a common estimation method used for categorical response data is weighted least square (WLS; Browne, 1984). This approach was developed for estimating a weight matrix based on an asymptotic covariance matrix that are typically estimated based on a non-Pearson correlation such as a tetrachoric correlation for dichotomous indicators or a polychoric correlation for polytomous indicators. The WLS fitting function is defined as

$$F_{wls} = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \quad (2.19)$$

Where  $\mathbf{s}$  is a vector containing unique elements of a sample correlation matrix (i.e., tetrachoric/polychoric correlation),  $\boldsymbol{\sigma}$  is a vector containing unique elements of the model-implied correlation matrix, and  $\mathbf{W}$  is a positive-definite weight matrix which is based on the variance and covariance of each element of  $\mathbf{s}$ .

If the weight matrix is correctly estimated, then the fitting function can lead to asymptotically efficient parameter estimates and correct standard errors and chi-square test statistic (Browne, 1984). However, this estimation method has been criticized due to its unstable performance in small or moderate samples. The weight matrix  $\mathbf{W}$  grows rapidly as the number of variables increases, and therefore can be extremely large when there are many indicators in a model. For example, consider a single factor model with 6 items ( $p = 6$ ). There are 21 elements in  $\mathbf{s}$ , that is  $v = 6(6+1)/2 = 21$ . Thus,  $\mathbf{W}$  is of the order  $v \times v$  ( $21 \times 21$ ) and has 231 distinct elements (i.e.,  $v(v+1)/2 = 21(22)/2 = 231$ ). In a slightly more realistic case, say, with 20 items ( $p = 20$ ),  $v = 210$  and the weight matrix includes 22155 unique elements. Unless the sample size is

sufficiently large,  $W$  is often nonpositive definite and cannot be inverted for a model with many indicators. Furthermore, a large sample size is also required for stable estimates of these asymptotic values in the weight matrix (Flora & Curran, 2004). Many simulation studies have shown that the WLS estimation produced a significant amount of bias in both estimated standard errors and test statistics (e.g., Dolan, 1994; Potthast, 1993). Consequently, the application of WLS is limited when the sample size is not sufficiently large.

To address the problems encountered with WLS, Muthén, du Toit, and Spisic (1997) introduced mean adjusted weighted least square (WLSM) as well as mean-and-variance adjusted weighted least square (WLSMV), both of which have been implemented in Mplus (Muthén & Muthén, 1998-2010). This adjusted WLS approach uses the diagonal elements of the weight matrix to estimate parameters so that the inverting of the weight matrix is avoided and computational burden is reduced. In order to correct the bias that results from the use of diagonal elements of the weight matrix instead of the full matrix, adjustments have been made for the standard errors of estimated parameters and the chi-square test statistics, which are similar to Satorra-Bentler adjustments (i.e., robust standard error and chi-square test) (Satorra & Bentler, 1994). In addition to the chi-square statistic, the model degrees of freedom is also adjusted in WLSMV. In their simulation study, Muthén et al (1997) found that WLSMV estimation was acceptable even for a sample size of 200. Some other simulation studies have confirmed this result by showing that WLSMV produced accurate test statistics, parameter estimates, and standard error estimates under a variety of conditions (e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004).

### **2.4.3 Previous Studies on the Comparisons of Estimation Methods across IRT and CFA**

By virtue of the analytical relationship between CFA and IRT, it is possible to examine performances of the estimation methods across the two frameworks. Knol and Berger (1991) used a simulation approach to compare NOHARM, TESTFACT, and a variety of common factor analysis approaches (principle factor analysis in SPSS, ML, Generalized Least Square) based on tetrachoric correlations in the recovery of MIRT parameters. The data were generated based on a two-parameter normal-ogive MIRT model with varied conditions including sample size (250, 500, 1,000), number of dimensions (1, 2, 3), and number of items (15, 30). They concluded that factor analysis on tetrachoric correlations performed at least as well as NOHARM or TESTFACT, although NOHARM and TESTFACT might be viewed as more theoretically appropriate for multidimensional data because they were developed specifically for MIRT.

Using both real and simulated data, Tate (2003) compared a number of methods including ULS using NOHARM, adjusted WLS using Mplus, MML/EM using TESTFACT, and some nonparametric methods. A number of datasets were simulated to reflect a variety of conditions in the IRT context such as unidimensionality, different patterns of item parameters, and different structures of multidimensionality. Each generated dataset was analyzed with estimation methods investigated in the study. With respect to parameter recovery, the results showed that the adjusted WLS approach performed fairly well except for the case in which guessing parameters were present in the data. The NOHARM and TESTFACT, on the other hand, recovered the item parameters well under most conditions including the conditions with guessing parameters. Similar results had been found by Finch (2010) when he examined the parameter estimation of MIRT models using NOHARM and adjusted WLS across a range of conditions such as sample size (250, 500, 1000, 2000), number of items (15, 30, 60), distribution

of latent traits (standard normal vs. nonnormal), and correlation between latent traits (0, 0.3, 0.5, 0.8). Instead of using qualitative measure such as poor, fair, and good for item recovery as Tate (2003) did, he computed RMSE to quantify the accuracy of parameter recovery. The results showed larger inaccuracy with the adjusted WLS estimation for item parameters when guessing parameters were present in the data. Otherwise, the estimation quality of the two approaches was quite comparable, particularly for the discrimination parameter. Furthermore, both methods were influenced by the distribution of a latent trait, and larger standard errors of item parameters were associated with skewed distributions.

In addition to the parameter estimation, the performance of Mplus, TESTFACT and NOHARM has been examined in testing dimensionality and a factor structure since all three methods can perform factor analysis. The comparisons, however, have been primarily made in the case of EFA since TESTFACT can only perform CFA for a specific structure (i.e., bi-factor model). The three estimation methods were found to perform reasonably well and result in similar solutions to identifying number of factors when the guessing parameter was not present in data (Stone & Yeh, 2006; Tate, 2003). On the other hand, when the guessing parameter was present in data, the adjusted WLS, which does not incorporate guessing, resulted in incorrect conclusion about dimensionality (Tate, 2003) while NOHARM and TESTFACT performed similarly well (Stone & Yeh, 2006; Tate, 2003).

Wirth and Edwards (2007) conducted a simulation study to compare MCMC for IRT models to the CFA estimation methods including WLS and adjusted WLS in terms of parameter recovery. Instead of simulating data based on IRT models, they simulated data using CFA models. The estimated IRT parameters were converted and compared to the corresponding CFA parameters. They generated data using a four-factor CFA model with 10 items loading on each

factor. The model included a mixture of item types with polytomous responses loading on the first factor and dichotomous responses loading on the other three factors. The generated data were analyzed using WLS, adjusted WLS and MCMC methods. The results showed that the adjusted WLS and MCMC methods produced estimates closer to the population values than the WLS method while MCMC provided estimates that were slightly closer to the population values than the adjusted WLS method. Their further analysis revealed that WLS produced some inappropriate values for the factor loadings, suggesting that the sample size used in the study might not be large enough to obtain a stable WLS solution. Based on the observed differences caused by different estimation methods, the authors recommended using various methods for the parameter estimation in order to achieve greater confidence in the final decision.

Forero and Maydeu-Olivares (2009) provided a comprehensive review of previous research on the empirical behavior of estimation methods within or across IRT and CFA frameworks. In addition, they conducted a simulation study that compared the performance of MML/EM in IRT and ULS in CFA in estimating a GR model under varied conditions (sample size, number of dimensions, size of factor loadings, number of items per factor, number of response categories per item, and item skewness). They concluded that the two methods were comparable with regard to the accuracy of parameter estimates and standard errors, but MML performed better in some harsh conditions such as a small number of indicators per dimension, item with high skewness, and small sample size.

Most previous comparative studies have examined MML and/or ULS for IRT models in comparison with WLS or adjusted WLS for CFA models. There have been few studies except for Wirth and Edwards (2007) that compared MCMC estimation in the IRT context to the CFA estimation methods. Even though Wirth and Edwards (2007) used a simulation approach to

compare MCMC and adjusted WLS, their study only involved one simulation condition, i.e., sample size, and only one replication was implemented for each condition. Furthermore, as most other studies, their study also used dichotomous models and was confined to a first-order factor. Consequently, with the rise of MCMC in the IRT framework, benefits of a higher-order model structure, and important application of polytomous items in real testing situations, this study primarily investigated the performance of MCMC for higher-order polytomous IRT models compared to WLSMV for higher-order CFA models. Through such a study, it is hoped to further explore the strengths and weaknesses of each methodology and guide researchers and practitioners to choose from or to integrate the two methodologies.



### **3.0 METHOD**

The primary purpose of this study was to compare the performance of CFA and IRT models with a higher-order structure in the parameter recovery of polytomous response data. In order to achieve this goal, a simulation study was conducted. In this chapter, the models used in the simulation study are presented first, then the estimation methods for each model are presented, and finally the simulation study including study design, data generation and estimation, and evaluation measures is described.

### **3.1 MODEL SPECIFICATION**

#### **3.1.1 Higher-order Graded Response IRT Model**

In a higher-order IRT model, the first-order model is structured with each item being loaded on one ability dimension, i.e., between-item multidimensionality. The correlation among domain abilities is explained by a general ability. In this study, this higher-order IRT model was extended to incorporate polytomous items. At the first order, GR model (Samejima, 1969) was used to model examinees' responses at domain levels. The probability of examinee  $j$  scoring  $x$  or above on item  $i$  of domain  $d$  is given by

$$P_{ijx(d)}^* = P(X_{ij} = x | \theta_{j(d)}, a_{i(d)}, b_{ix(d)}) = \frac{\exp[Da_{i(d)}(\theta_{j(d)} - b_{ix(d)})]}{1 + \exp[Da_{i(d)}(\theta_{j(d)} - b_{ix(d)})]} \quad (3.1)$$

where

$X_{ij}$  is the response of examinee  $j$  to item  $i$ ;

$\theta_{j(d)}$  is the ability for examinees  $j$ ;

$a_{i(d)}$  is the slope/discrimination parameter for item  $i$ ;

$b_{i1(d)} \dots, b_{im_i(d)}$  are the threshold parameters of item  $i$ ;

$D=1$  is used in the current study;

$x = 0, \dots, m_i$  number of score values for item  $i$ ;

$j = 1, 2, \dots, N$  number of examinees;

$i = 1, 2, \dots, I$  number of items;

$d = 1, 2, \dots, D$  number of domains or dimensions;

The probability of examinee  $j$  earning a particular score category of item  $i$  in domain  $d$  is given by

$$\begin{cases} P_{ij0(d)} = 1 - P_{ij1(d)}^* \\ P_{ijx(d)} = P_{ijx(d)}^* - P_{ij(x+1)(d)}^* \quad x = (1, 2, \dots, m_i - 1) \\ P_{ijm_i(d)} = P_{ijm_i(d)}^* - 0 \end{cases} \quad (3.2)$$

This formulation of GR model incorporates multiple ability dimensions consisting of independent clusters of items. Each domain test can be regarded as a unidimensional IRT model but correlated to other domains when estimated. Meanwhile, other forms of polytomous response model such as generalized partial credit models can be used in replace of GR model.

If let the parameters for the  $i$ th item in the domain  $d$  be  $\beta_{i(d)} = (a_{i(d)}, b_{i1(d)}, \dots, b_{im_i(d)})$ , the likelihood of an item response in a domain test  $X_{ij(d)}$  is denoted as

$$L_{ij(d)} = L(X_{ij(d)}|\theta_{j(d)}, \beta_{i(d)}) = \prod_{x=0}^{m_i} (P_{ijx(d)})^{X_{ijx(d)}} \quad (3.3)$$

where  $X_{ijx(d)}=1$  if  $X_{ij(d)} = x + 1$ ;  $X_{ijx(d)}=0$  otherwise,

Then the likelihood function of the response matrix  $\mathbf{X}$  is

$$L(\mathbf{X}|\boldsymbol{\theta}_\delta, \boldsymbol{\beta}) = \prod_{j=1}^N \prod_{d=1}^D \prod_{i(d)=1}^{I(d)} (L_{ij(d)}) \quad (3.4)$$

where  $\boldsymbol{\theta}_\delta = (\theta_{j(d)}); \boldsymbol{\beta} = (\beta_{i(d)})$

At the second order of the model, the domain abilities are related to the overall ability via a linear function  $\theta_{j(d)} = \gamma_{(d)}\theta_j + \varepsilon_{j(d)}$ , where  $\gamma_{(d)}$  is the latent regression coefficient of the domain ability  $\theta_{j(d)}$  on the overall ability  $\theta_j$ , and  $\varepsilon_{j(d)}$  is the error term that is independent of other error terms and follows a normal distribution with mean of zero and variance of  $1 - \gamma_{(d)}^2$ . It is assumed that  $\theta_j \sim N(0,1)$ . Conditional on the regression coefficient and overall ability, the domain ability follows  $\theta_{j(d)}|\gamma_{(d)}, \theta_j \sim N(\gamma_{(d)}\theta_j, 1 - \gamma_{(d)}^2)$ . The marginal distribution of each domain ability is also standard normal (i.e.,  $\theta_{j(d)} \sim N(0,1)$ ).

### 3.1.2 Second-order CFA Model

The structure of the second-order CFA model is the same as that of HO-IRT model, but the model is formulated in a different way. At the first order, the unobserved continuous variables that underlie each item with categorical responses are modeled as follows,

$$\mathbf{y} = \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (3.5)$$

where  $\mathbf{y}$  is a  $p \times 1$  vector of unobserved continuous variables;

$\mathbf{A}$  is a  $p \times d$  matrix of factor loadings;

$\boldsymbol{\eta}$  is a  $d \times 1$  vector of latent variables/factors;

$\boldsymbol{\varepsilon}$  is a  $p \times 1$  vector of unique variances of  $\mathbf{y}$ .

With a simple structure at the first order, the model equation can be expressed as,

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ \vdots \\ y_{p-2} \\ y_{p-1} \\ y_p \end{pmatrix} = \begin{pmatrix} \lambda_{y11} & 0 & & 0 \\ \lambda_{y21} & 0 & \dots & 0 \\ \lambda_{y31} & 0 & & 0 \\ 0 & \lambda_{y42} & & 0 \\ 0 & \lambda_{y52} & \dots & 0 \\ 0 & \lambda_{y62} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_{y(p-2)d} \\ 0 & 0 & & \lambda_{y(p-1)d} \\ 0 & 0 & & \lambda_{ypd} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \vdots \\ \varepsilon_{p-2} \\ \varepsilon_{p-1} \\ \varepsilon_p \end{pmatrix} \quad (3.6)$$

At the second order, the relation among  $\boldsymbol{\eta}$  and second order factor  $\boldsymbol{\xi}$  is defined as

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (3.7)$$

where  $\boldsymbol{\xi}$  is a  $n \times 1$  vector of 2<sup>nd</sup> factors;

$\boldsymbol{\Gamma}$  is a  $d \times n$  matrix of factor loadings of 1<sup>st</sup> order factors on 2<sup>nd</sup> order;

$\boldsymbol{\zeta}$  is a  $d \times 1$  vector of unique variances of  $\boldsymbol{\eta}$ .

If only one higher-order factor is defined, the model equation is

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_d \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_d \end{pmatrix} \xi + \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_d \end{pmatrix} \quad (3.8)$$

The covariance of  $\boldsymbol{\zeta}$ , i.e., the unique variance at the second order is

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & 0 & \dots & 0 \\ 0 & \psi_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_{dd} \end{pmatrix} \quad (3.9)$$

where  $\psi_{dd} = 1 - \gamma_{(d)}^2$ .

## 3.2 SIMULATION STUDY

### 3.2.1 Study Design

A 2 x 2 x 2 x 2 x 2 mixed design simulation study was conducted to evaluate the performance of two estimation methods (MCMC in HO-IRT vs. WLSMV in second-order CFA) under various conditions. In addition to the estimation method as a within-subjects independent variable (IV), four between-subjects IVs were manipulated to represent different test situations. These between-subjects IVs include correlations between domain abilities, number of examinees, number of domains, and number of items in each domain. Table 3.1 lists all the IVs and their levels specified in the current study.

Table 3.1 Simulated IVs and Levels for Each IV

Estimation methods	Correlation between domains (in terms of $\gamma_{(d)}$ )	Number of examinees ( $N$ )	Number of domains ( $D$ )	Number of items in each domain ( $J$ )
MCMC	Moderate (0.4-0.6)	500	3	5
WLSMV	High (0.6-0.8)	1000	5	10

Note: MCMC = Markov chain Monte Carlo; WLSMV = mean-and-variance adjusted weighted least square

Two levels (moderate vs. high) of varied correlation between domain abilities were specified in terms of the regression coefficients of domain abilities on the overall ability ( $\gamma_{(d)}$ ). The moderate correlation level was manipulated to vary from 0.4 to 0.6. This corresponds to the correlations between domain abilities ( $\rho$ ) varying from 0.16 to 0.36 in that the correlation between two domain abilities is computed as the product of corresponding regression coefficients. The higher level of correlation was specified to vary from 0.6 to 0.8 (corresponding to  $\rho$  varying from 0.36 to 0.64). Previous studies examining the performance of MIRT with

MCMC estimation or CFA with adjusted WLS have included a series of correlation levels between dimensions ( $\rho = 0$  to  $0.9$ ) (e.g., de la Torre, 2009; Finch, 2010; Yao, 2010; Yao & Boughton, 2007). Such a wide range of correlation values have also been used to assess HO-IRT models. For example, Sheng and Wikle (2008) used values of 0, 0.3, 0.6, 0.8, and 1 while de la Torre and Song (2009) used  $\rho = 0, 0.4, 0.7, 0.9$ . In most of these studies, no correlation between domain abilities (i.e.,  $\rho = 0$ ) was used to simulate the situation where a unidimensional IRT calibration for each dimension was appropriate. On the other hand, when the dimensions were highly correlated (e.g.,  $> 0.8$ ), the differences between the dimensions became trivial, i.e., unidimensionality across all items could be assumed. Therefore, the values chosen in the current study were to simulate the situations where the use of MIRT/HO-IRT models is most appropriate. In addition, it might be unreasonable to assume an equal degree of correlations among dimensions that was used by most previous studies (e.g., de la Torre, 2009; Finch, 2010). Thus, in the current study, the correlations among dimensions were manipulated to vary within one specified correlation condition.

Adequacy of sample size has been one of important issues that were studied and discussed in the IRT or CFA analysis. Reise and Yu (1990) recommended at least 500 examinees were needed for adequate calibration of a graded response IRT model. In their simulation study that compared MCMC and MML in the recovery of graded response model parameters, Kieftenbeld and Natesan (2012) indicated that a sample size of 300 seems to be a necessary minimum, but 500 is better. Some simulation studies in linking and equating that included item calibration components have also provided some guidelines for sample size in the GR model estimation. For example, in the equating studies conducted by Cohen and Kim (1998) and Kim and Cohen (2002), they used sample sizes of 300 and 1000. The results suggested that the sample

size of 1000 produced better and adequate item parameter estimates. Meanwhile, with the generalized partial credit models, the sample size of 1000 has been used in many simulation studies examining the performance of MIRT with MCMC estimation (e.g., de la Torre, 2009; Hong, et al., 2010). Thus, the studies of polytomous IRT models indicate a sample size of 1000 being more than adequate for the GR model estimation. In the framework of CFA, previous simulation studies using adjusted WLS have examined a range of values for sample size from 100 (Flora & Curran, 2004) to 2000 (Finch, 2010; Tate, 2003). Taken together the information provided by previous research in both IRT and CFA frameworks, the number of examinee in this study was simulated to be 500 and 1000 so as to represent relatively small to fairly large sample size conditions for both types of measurement models.

The number of dimensions was simulated to be 3 and 5. These two levels were chosen to represent the number of subscales or sections that have been commonly used in many operational or simulation studies. For example, Sinharay (2010) reviewed several operational datasets in education and found the number of dimensions ranged from 2 to 7 with several having three dimensions. Meanwhile, the condition of five-dimension test has been included in many previous simulation studies for MIRT or HO-IRT (e.g., de la Torre & Patz, 2005; de la Torre & Song, 2009). This condition was, therefore, included in the current study for the comparison purpose. Although most previous studies in MIRT or HO-IRT also simulated two dimensions (e.g. de la Torre & Hong, 2010; de la Torre & Song, 2009; Sheng & Wikle, 2008), this level was not included in the current study in that the CFA model with only two factors at the first level might have an identification problem for the second-order factor. The number of items in each dimension was specified at 5 and 10. These values are adequate and appropriate due to the polytomous nature of items. Furthermore, combining the levels for the number of items in each

domain with varied number of dimensions, the test lengths ranged from 15 to 50, which represents short to relatively long tests.

The levels of between-subjects IVs specified in the present study resulted in 16 ( $2 \times 2 \times 2 \times 2 = 16$ ) combinations of conditions. Under each condition, the steps for data generation and analysis were described as following:

- 1) Based on the simulating item and ability parameters (described in the section 3.2.2), the item responses were generated under the HO-IRT GR model described in the section 3.1.1.
- 2) The generated data were analyzed with the HO-IRT model using MCMC in WinBUGS 1.4 (Spiegelhalter, et al, 2003) and the second-order CFA model using WLSMV in Mplus Version 6. (Muthén & Muthén, 1998-2010).
- 3) The estimated item parameters and overall and domain abilities were compared to the corresponding simulating parameter values in terms of evaluation measures (explained in the section 3.4).
- 4) The steps 1 to 3 are repeated 30 times for each condition.

### **3.2.2 Simulation Procedure**

First, the item parameters under the GR model were simulated. The discrimination parameters ( $a$ ) were specified to be randomly varied across dimensions with three levels of uniform distributions:  $U(1.0, 1.5)$ ,  $U(1.2, 2.3)$ , and  $U(2.0, 2.5)$ . These three distributions have means of 1.25, 1.75, and 2.25 respectively to reflect moderately low, average, and moderately high item discrimination parameters that are commonly seen in many educational and psychological measures. The discrimination parameter plays a very important role in determining the



information provided by items. Generally speaking, items with high discrimination provide more information but over a narrow range while less discriminating items provide less information but over a wider range. Furthermore, information is often described as IRT version of reliability since it is closely related to measurement errors. In general, the higher the information, the lower the errors. Thus, the present study used varied levels of discrimination parameters across subscales to reflect the varied information or loosely speaking, reliability, provided by each subscale. Specifically, for the three-dimension condition, the item discrimination parameters of each subscale were generated from the three uniform distributions respectively. For the five-dimension condition, the first three dimensions were generated from the three uniform distributions respectively, and the last two were generated from  $U(1.0, 1.5)$  and  $U(1.2, 2.3)$  respectively.

Next, the item threshold values were simulated. The number of response categories of each item was fixed at 5, and therefore 4 threshold parameters were generated for each item. To represent a varied range of items involved in a test, for each five-category item, the first threshold was generated from a  $U(-2, -1)$  distribution, and then the subsequent threshold parameters were obtained by adding to its previous threshold value a constant randomly selected from 0.75 or 1. For example, if the first threshold of an item was -1, the other three threshold values for this item could be either -0.25, 0.5, and 1.25 (by adding 0.75), or 0, 1, and 2 (by adding 1). Such configurations of the threshold parameters were intended to reflect not only a wide range of trait levels (-2~2) covered by the items but also variability of the thresholds across items.

To provide a general idea on the breadth of ability measurement provided by each subscale, the information function of each subscale under the condition of three-dimension with

five-item on each dimension was plotted based on one set of simulating item parameters. This set of item parameters was provided in Table 3.2, and the subscale information was plotted in Figure 3.1. As Figure 3.1 shows, the higher the discrimination parameters, the more information the subscale provided. Thus, the third subscale provided the most information. Meanwhile, within each subscale, the maximum information was in general related to the ability across the range of -1 to 1.

Table 3.2 An Example Set of Simulated Item Parameters

Subscales	$a$	$b_1$	$b_2$	$b_3$	$b_4$
Subscale 1	1.20	-1.58	-0.83	-0.08	0.67
	1.18	-1.99	-0.99	0.01	1.01
	1.33	-1.59	-0.84	-0.09	0.66
	1.20	-1.86	-0.86	0.14	1.14
	1.47	-1.82	-1.07	-0.32	0.43
Subscale 2	1.64	-1.28	-0.28	0.72	1.72
	1.60	-1.65	-0.90	-0.15	0.60
	1.92	-1.38	-0.63	0.12	0.87
	1.64	-1.45	-0.45	0.55	1.55
	2.23	-1.88	-0.88	0.12	1.12
Subscale 3	2.20	-1.52	-0.77	-0.02	0.73
	2.18	-1.84	-0.84	0.16	1.16
	2.33	-1.28	-0.53	0.22	0.97
	2.20	-1.93	-0.93	0.07	1.07
	2.47	-1.19	-0.19	0.81	1.81

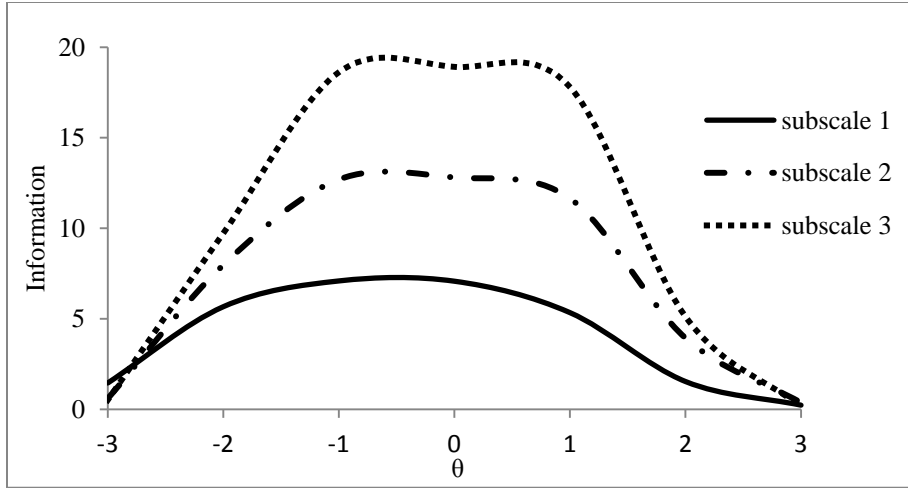


Figure 3.1 Subscale information for a three-dimension test with five items in each dimension

After the item parameters were generated, the examinees' ability on both overall and domain levels were simulated. For a specific sample size, the overall ability was generated from  $N(0, 1)$ . For each specified level of domain correlations, the regression coefficients of domain abilities on the overall ability ( $\gamma_{(d)}$ ) was generated from the specified uniform distribution, i.e.,  $\gamma_{(d)} \sim U(0.4, 0.6)$  or  $\sim U(0.6, 0.8)$ . The error terms ( $\varepsilon_{j(d)}$ ) were generated from a normal distribution with a mean of 0 and variance of  $1 - \gamma_{(d)}^2$ . Given the overall ability, generated regression coefficients, and error terms, the domain abilities were computed as  $\theta_{j(d)} = \gamma_{(d)}\theta_j + \varepsilon_{j(d)}$ . Based on the abilities and items in each domain, the probability of responding in a particular category was computed using the GR model. Then the examinees' responses were generated by comparing the cumulative probabilities of the response categories to the random numbers generated from a standard uniform distribution. This data generation procedure was implemented using SAS/IML programming. Appendix A provides the SAS code used to generate higher-order GR data under the condition of three domain abilities and five items in each domain. 30 sets of response data were generated under each condition specified in the current study.

### 3.2.3 Validation of Data Generation

Data validation is a process to check if the generated data meet certain criteria intended by the researchers or simulation models. It is necessary because incorrectly generated data could lead to biased even distorted results. Although various conditions have been involved in the current study, the logic of data generation process is identical. Therefore, the data validation was done using one dataset that was generated under the condition involving three moderately correlated domains with each five items distinctively measuring one domain. To verify the process of generating item responses, the observed and model-based proportions of examinees responding in each category were compared. A small difference between the two proportions is an indication of appropriate data generation process. A sufficiently large sample size of 10,000 examinees was simulated. Due to the between-item multidimensionality structure at the first order of the HO-IRT model, each set of five items were generated using a distinctive domain ability distribution. However, the way in which each set of five items was generated was identical. Therefore, the observed and model-based proportions of examinees responding in each category were only calculated for the first five items in the first dimension by fixing the domain ability ( $\theta_i$ ) at 0. As shown in Table 3.3, the absolute differences between the two proportions were all equal to or less than 0.01, which indicates the data were correctly generated.

Table 3.3 Observed and Expected Proportions of Response Categories

Items	Observed Proportions					Expected Proportions				
	Cat1	Cat2	Cat3	Cat4	Cat5	Cat1	Cat2	Cat3	Cat4	Cat5
1	0.14	0.14	0.19	0.21	0.33	0.15	0.14	0.19	0.20	0.32
2	0.06	0.15	0.29	0.30	0.20	0.06	0.14	0.30	0.30	0.20
3	0.10	0.20	0.34	0.23	0.12	0.10	0.21	0.33	0.24	0.12
4	0.09	0.17	0.29	0.27	0.19	0.08	0.17	0.30	0.27	0.19
5	0.09	0.12	0.19	0.23	0.37	0.09	0.11	0.19	0.23	0.37

An alternative way to validate the data generation process is to examine if the structure of generated data conforms to the intended model structure. In a HO-IRT model, between-item multidimensionality is assumed at the first order. To verify this structure in the generated response data, an EFA was performed using the WLSMV method in Mplus. With three factors extracted, the factor pattern with promax rotation was shown in Table 3.4. From this table, one can easily see that each item was loaded on only one single factor, and each set of five items was loaded on one distinctive factor. This simple three-factor structure was exactly the same as the first-order model structure of the generated data.

Table 3.4 Factor Pattern for Exploratory Factor Analysis with Promax Rotation

Items	F1	F2	F3
1	<b>0.57</b>	0.01	0.00
2	<b>0.56</b>	-0.01	0.01
3	<b>0.62</b>	0.01	0.00
4	<b>0.56</b>	0.02	0.01
5	<b>0.66</b>	-0.02	-0.02
6	0.01	0.01	<b>0.68</b>
7	0.01	-0.01	<b>0.68</b>
8	0.00	0.00	<b>0.74</b>
9	0.00	0.00	<b>0.68</b>
10	-0.01	0.01	<b>0.79</b>
11	0.01	<b>0.77</b>	0.00
12	0.00	<b>0.78</b>	0.00
13	0.00	<b>0.79</b>	0.01
14	-0.01	<b>0.78</b>	0.00
15	0.01	<b>0.80</b>	0.00

As EFA cannot be used to confirm the higher-order factor structure, a CFA was performed on the same dataset using the WLSMV method in Mplus. Both the chi-square test and fit indices indicated good model fit ( $\chi^2(87) = 81.14$ ,  $p = .66$ ; RMSEA = .000, 90% CI [.000~.005]; WRMR = .55). Furthermore, a good recovery of regression coefficients ( $\gamma_{(d)}$ ) is also an indication of correct data generation. As shown in Table 3.5, the differences between the

simulated and estimated  $\gamma_{(d)}$  values were less than 0.01. Based on all the evidence, the data generation process was appropriate.

Table 3.5 Recovery of Regression Coefficients of Domain Abilities on Overall Ability

Regression Coefficients ( $\gamma_{(d)}$ )	True	Estimated	Difference (Estimated -True)
$\gamma_1$	0.54	0.55	.01
$\gamma_2$	0.55	0.55	.00
$\gamma_3$	0.47	0.48	.01

### 3.3 ESTIMATION

Each generated dataset was analyzed with the HO-IRT model using MCMC in WinBUGS and the second-order CFA model using WLSMV in Mplus. With the MCMC estimation method, the prior distributions were specified in WinBUGS as follows.

For the ability estimation:

$$\theta_j \sim N(0, 1);$$

$$\gamma_{(d)} \sim U(0, 1);$$

$$\varepsilon_{j(d)} \sim N(0, 1 - \gamma_{(d)}^2); \text{ and } \theta_{j(d)} = \gamma_{(d)}\theta_j + \varepsilon_{j(d)} \text{ can be estimated accordingly.}$$

For the item parameters:

$$a_{i(d)} \sim U(1, 2.5);$$

$$b_{i1(d)} \sim U(-3, -1);$$

$$b_{i(x+1)(d)} \sim U(b_{ix(d)}, 3) \text{ where } x = 1, 2 \text{ or } 3.$$

As reviewed in Chapter II, the most critical step in MCMC is to assess if the Markov chains have converged. Therefore, a preliminary analysis was performed to determine how many

initial iterations should be discarded (burn-in rate) and whether sufficient iterations have been run to summarize the parameters' posterior distributions. The response data were simulated for 500 examinees responding to 15 polytomous items with each set of five items measuring one single domain. The WinBUGS code for analyzing this preliminary dataset was shown in Appendix B. The regression coefficients of domain abilities on the overall ability were generated from  $U(0.4, 0.6)$ . In this preliminary study, three chains with randomly generated initials were run. If the chains from different starting points result in the same stationary distribution, there will be a strong likelihood of convergence. The burn-in rate for all the chains was set at 5000. Each chain was run for another 7500 iterations and thinned by keeping every 5<sup>th</sup> iteration to reduce any autocorrelation. The convergence of the item parameter estimates was examined by a visual inspection of several diagnostic plots available in WinBUGS. First, the history plots showing the trajectories of the samples of item parameters produced by all three chains were inspected. Figure 3.2 shows the history of one discrimination and four threshold parameters for the first item. In each plot, the three chains, started at dispersed initial values, have consistently merged well and each stably moved around a particular space, indicating the chains have converged to a common distribution. Furthermore, after 5000 iterations, the trajectories remain stable, which means both the burn-in rate and number of additional iterations were sufficient. Similar patterns of history plots have been found for parameter estimates of other items.

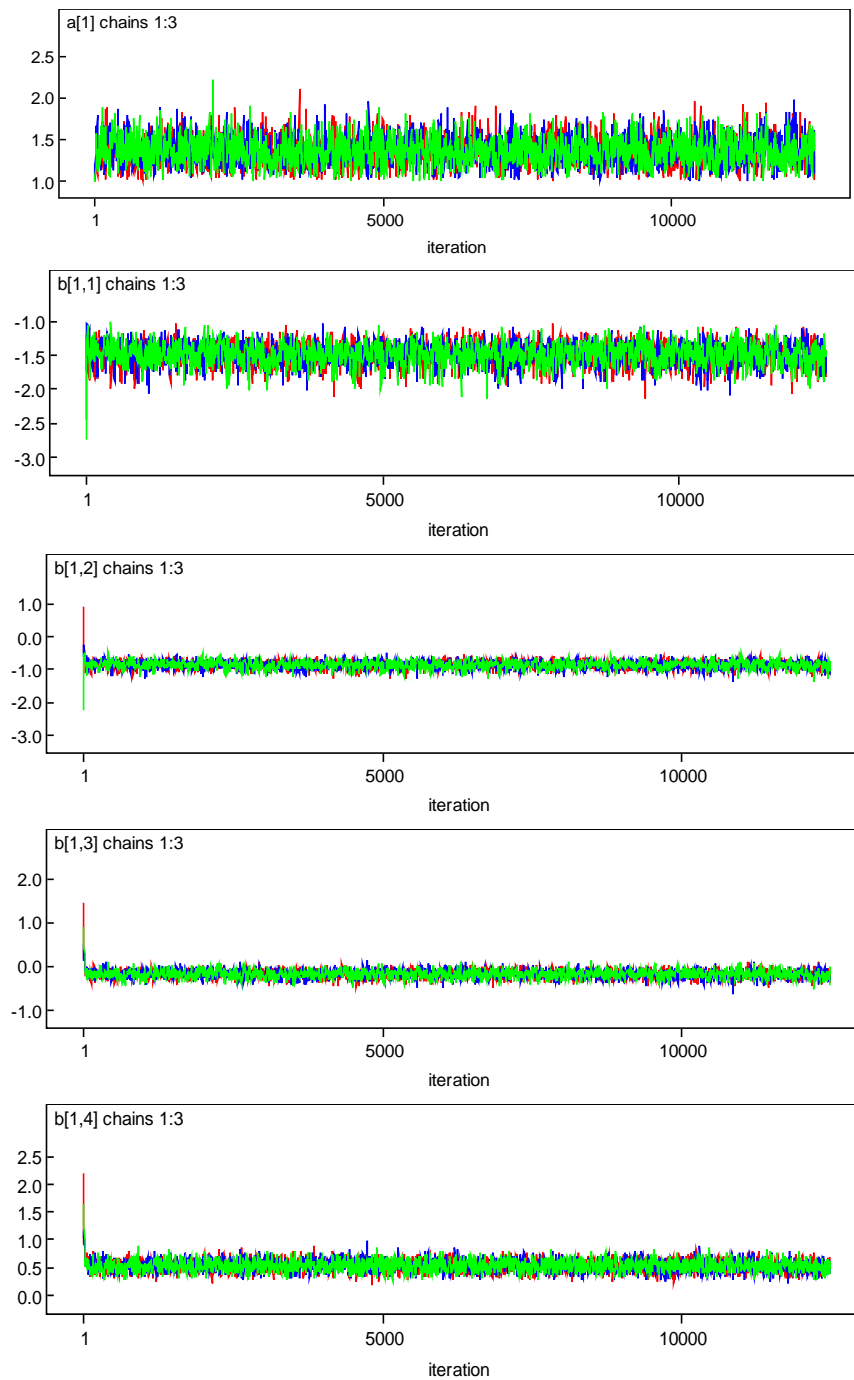


Figure 3.2 History plots of discrimination and threshold parameters for item 1



Secondly, the ‘BGR diagnostic’ plots were used to assess the convergence of multiple chains. In this plot, the red line represents the Gelman-Rubin statistic  $\hat{R}$ , which compares the ratio of the pooled chain variance to the within-chain variance. The green and blue lines represent respectively the width of 80% interval of the pooled runs and average width of 80% intervals within the individual runs. Convergence may be achieved when the red line is close to 1, and the blue and green lines stabilize to some number. Figure 3.3 includes the BGR diagnostic plots for the discrimination and threshold parameter estimates of item 1. Based on these plots, the model converged after 5000 iterations: the red lines were close to 1 and the blue and green lines were almost overlapped for all the iterations after the burn-in rate. Similar results were found for parameter estimates of other items.

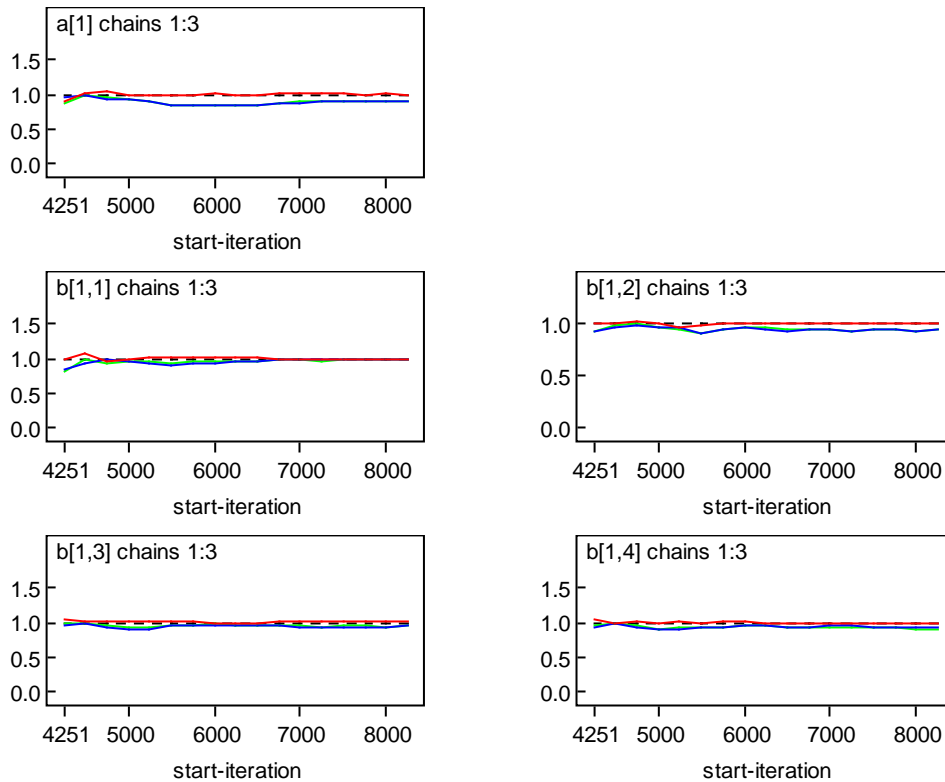


Figure 3.3 BGR diagnostic plots of discrimination and threshold parameters for item 1

Finally, the autocorrelation plots were examined. High autocorrelation can result in slow convergence and thus affect the efficiency of the MCMC simulation. As shown in Figure 3.4, close to zero correlation of the sequential draws of the parameters in each chain at later time lags implied the convergence of the chains.

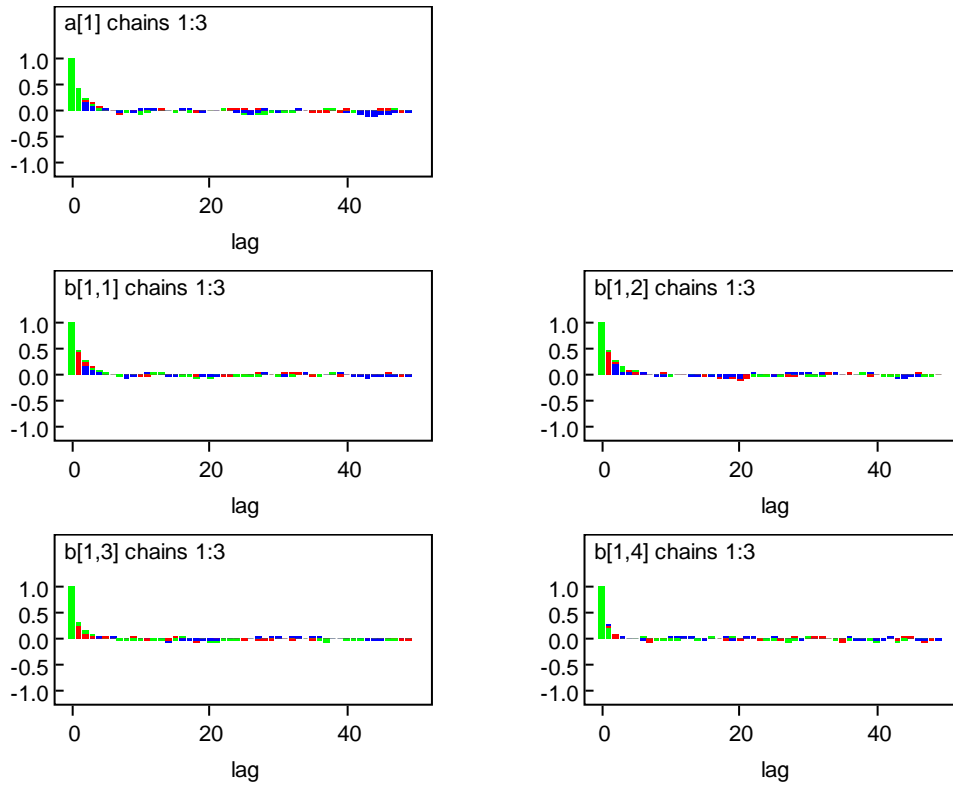


Figure 3.4 Autocorrelation plots of discrimination and threshold parameters for item 1

Meanwhile, a good recovery of item parameters is also an indication of estimation convergence. Therefore, the true and estimated item parameters were compared. As shown in Table 3.6, average absolute difference between the true and estimated values was 0.12 for the discrimination parameters and 0.08 for the threshold parameters, which were acceptable in terms of estimation accuracy.

Table 3.6 True and Estimated Parameter Values in WinBUGS

Dimension	Item	True						Estimated					
		a	b1	b2	b3	b4	$\gamma$	a	b1	b2	b3	b4	$\gamma$
D1	1	1.20	-1.58	-0.83	-0.08	0.67	0.54	1.36	-1.49	-0.87	-0.18	0.53	0.51
	2	1.18	-1.99	-0.99	0.01	1.01		1.16	-2.02	-1.17	-0.02	1.11	
	3	1.33	-1.59	-0.84	-0.09	0.66		1.15	-1.81	-1.05	-0.12	0.68	
	4	1.20	-1.86	-0.86	0.14	1.14		1.10	-1.87	-0.79	0.13	1.10	
	5	1.47	-1.82	-1.07	-0.32	0.43		1.43	-1.96	-1.04	-0.37	0.29	
D2	6	1.64	-1.28	-0.28	0.72	1.72	0.42	1.45	-1.38	-0.30	0.84	1.97	0.46
	7	1.60	-1.65	-0.90	-0.15	0.60		1.32	-2.09	-1.09	-0.11	0.70	
	8	1.92	-1.38	-0.63	0.12	0.87		1.91	-1.46	-0.66	0.12	0.87	
	9	1.64	-1.45	-0.45	0.55	1.55		1.75	-1.40	-0.35	0.52	1.60	
	10	2.23	-1.88	-0.88	0.12	1.12		2.01	-2.05	-1.02	0.22	1.23	
D3	11	2.20	-1.52	-0.77	-0.02	0.73	0.59	2.31	-1.50	-0.71	-0.06	0.77	0.64
	12	2.18	-1.84	-0.84	0.16	1.16		2.08	-1.92	-0.91	0.04	1.16	
	13	2.33	-1.28	-0.53	0.22	0.97		2.31	-1.38	-0.64	0.21	0.96	
	14	2.20	-1.93	-0.93	0.07	1.07		2.31	-1.82	-0.83	0.12	1.08	
	15	2.47	-1.19	-0.19	0.81	1.81		2.35	-1.21	-0.11	0.81	1.75	

In addition to the convergence of item parameter estimates, the convergence of the regression coefficients estimates ( $\gamma$ ) was also evaluated in that the domain ability estimates were closely related to the  $\gamma$  estimates in higher-order structured models. As for the item parameters, in a total of 12500 iterations for each chain, the first 5000 iterations were discarded, and the remaining 7500 iterations were thinned by taking every 5<sup>th</sup> iteration. The convergence was evaluated by a visual inspection of history plots, BGR diagnostic plots, and autocorrelation plots. Figures 3.5 and 3.6 show the history and autocorrelation plots for the three  $\gamma$  estimates, respectively. From these figures, one can see that all the estimates would most likely converge, particularly for the first two  $\gamma$ s. The histories were largely overlapped among multiple chains, and autocorrelations were quickly reduced to close to zero. Even though the third  $\gamma$  estimate seems not to have converged as well as the first two, it is still acceptable. A further examination of BGR diagnostic plots (Figure 3.7) also indicates that the estimates have converged after the first 5000 iterations. Furthermore, the recovery of the  $\gamma$  estimates was examined. As shown in Table 3.6, the absolute differences between the true and estimated values were 0.03, 0.04 and

0.05, which indicates that the regression coefficients were well estimated with the specified number of iterations.

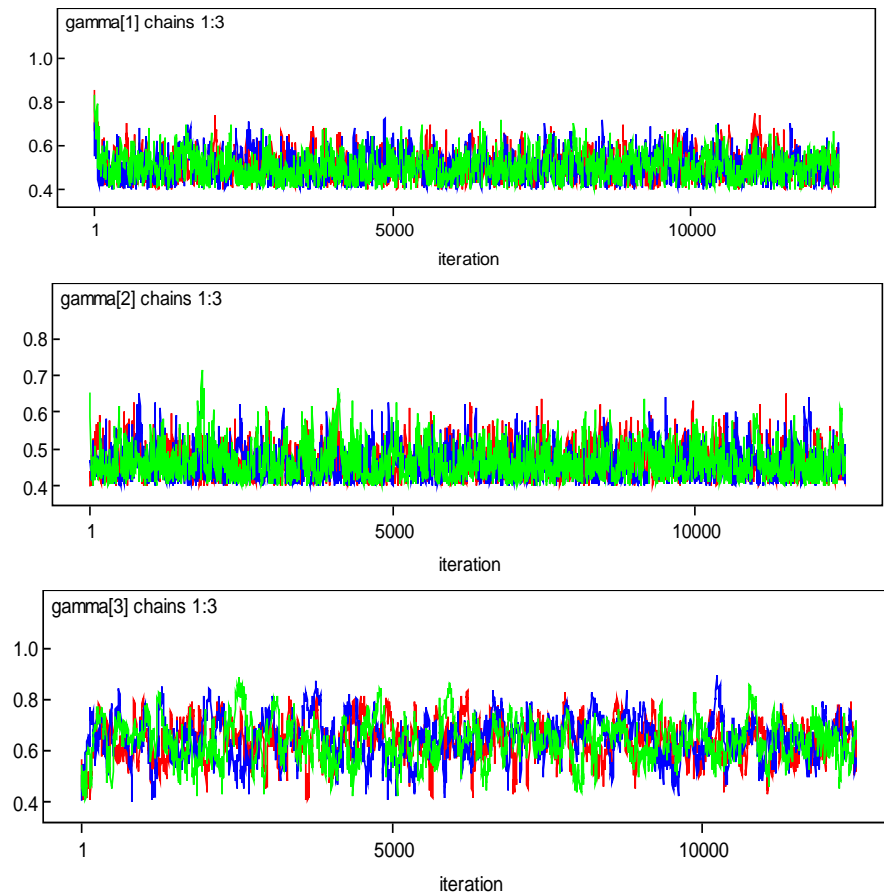


Figure 3.5 History plots of  $\gamma$  estimates for three domain abilities

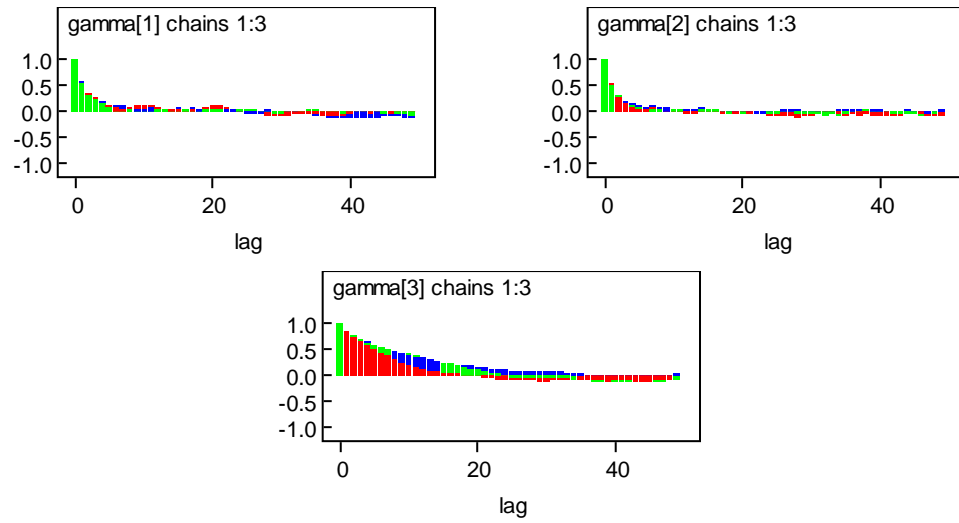


Figure 3.6 Autocorrelation plots of  $\gamma$  estimates for three domain abilities

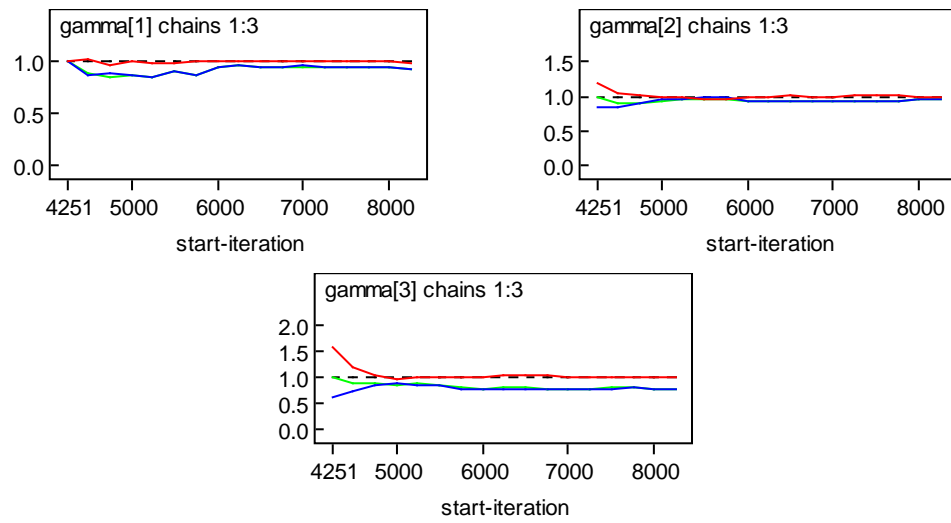


Figure 3.7 BRG plots of  $\gamma$  estimates for three domain abilities

For the second-order CFA model, the WLSMV estimation was implemented in Mplus. The basic WLS fitting function is defined as

$$F_{wls} = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \quad (3.10)$$

where  $\mathbf{s}$  is a vector containing unique elements of a sample correlation matrix (i.e., polychoric correlation);  $\boldsymbol{\sigma}$  is a vector containing unique elements of the model-implied correlation matrix; and  $\mathbf{W}$  is a positive-definite weight matrix which is based on the variance and covariance of each element of  $\mathbf{s}$ .

With the WLSMV approach, the diagonal elements of the weight matrix were used to estimate the parameters. In order for the factors to be identified in a CFA model, either factor variance or one of the factor loadings needs to be fixed to 1. For the second-order factor in the current study, its variance was fixed to 1 (i.e., standardized metric for higher-order factor). By doing so, however, the variance of the first order factor cannot be fixed to 1. Therefore, one of the factor loadings was fixed at 1. The standardized parameter estimates were requested for the final analysis. To evaluate if the model converged with the WLSMV estimation method, the same data used for MCMC estimation were analyzed in Mplus. Appendix C provides the Mplus code used for this analysis. With the default number of iterations (1000) and convergence criterion (0.0005), the results indicated that the model estimation converged.

### 3.4 EVALUATION CRITERIA

To compare the item parameter estimates between MCMC and WLSMV, the parameter estimates obtained from the second-order CFA model were converted to the corresponding IRT parameters using the following equations:

$$a_i = \frac{\lambda_i}{\sigma_i} D \quad (3.11)$$

$$b_{ix} = \frac{\tau_{ix}}{\lambda_i} (x = 1, \dots, m_i) \quad (3.12)$$

where  $a_i$  and  $b_{ix}$  represent the discrimination and threshold parameters respectively for item  $i$  from IRT models;  $\lambda_i$  and  $\tau_{ix}$  are the factor loading and threshold parameters respectively for item  $i$  from CFA models;  $\sigma_i = \sqrt{1 - \lambda_i^2}$ ; and  $D = 1.7$  scaling constant.

The accuracy of item parameter recovery was then evaluated by computing the root mean square error (RMSE) and bias. The RMSE and bias were defined as

$$RMSE = \sqrt{\left( \frac{1}{p} \sum_{i=1}^p (\hat{\varphi}_i - \varphi_i)^2 \right)} \quad (3.13)$$

$$BIAS = \frac{1}{p} \sum_{i=1}^p (\hat{\varphi}_i - \varphi_i) \quad (3.14)$$

where  $\hat{\varphi}$  is the estimated parameter obtained from the simulated data;  $\varphi$  is the true parameter value; and  $p$  is the number of parameters. If the estimated parameter values recover the true values well, lower RMSE is expected. The statistic BIAS can indicate the direction of the bias (over or under) in parameter estimation to some extent.

To evaluate the accuracy of recovering regression coefficients for both estimation methods, the RMSE of regression coefficients was computed. Furthermore, a 2 x 2 x 2 x 2 x 2 mixed analysis of variance (ANOVA) was performed on the RMSE of the item parameter estimates and on the RMSE of the regression coefficient estimates across the 30 replications respectively as a function of the estimation methods, number of examinees, number of domains,

number of items in each domain, and correlations between domain abilities. Although the analyses would result in higher-order interactions such as four-way and five-way interactions, the interpretation has been focused on no more than three-way effects (i.e., main effects, two-way and three-way interactions) because effects higher than three-way would be very difficult to interpret and usually have less significant meaning for practice. Meanwhile, based on Cohen's (1988) general guidelines of effect size in ANOVA (partial  $\eta^2$ : small: 0.01; medium: 0.06; large: 0.14), only effects (particularly for interaction effects) with moderate or large effect size were reported and further analyzed in the current study.

To compare the HO-IRT and second-order CFA models in the overall ability recovery, the correlation between the estimated abilities and true abilities were computed. Specifically, the second-order factor scores obtained from the second-order CFA model and the overall ability obtained from the HO-IRT model were compared to the true overall ability. The same statistics were computed for the domain abilities. Furthermore, a mixed ANOVA was performed on the correlation between the true and estimated overall ability as a function of all the conditions examined in the current study.



## 4.0 RESULTS

### 4.1 ITEM PARAMETER RECOVERY

To compare the recovery of item parameter estimates between the HO-IRT model and second-order CFA model, a mixed ANOVA was conducted on the RMSE of both item discrimination and threshold estimates. With the estimation method being a within-subjects variable, between-subjects IVs included number of domains, number of examinees, number of items in each domain, and correlation between domains.

For the item discrimination estimates, the MCMC method ( $M = 0.14$ ,  $SE = 0.002$ ) under the HO-IRT model produced slightly lower RMSE than the WLSMV method ( $M = 0.17$ ,  $SE = 0.002$ ) under the second-order CFA model,  $F(1, 464) = 683.64$ ,  $p < .001$ ,  $\eta_p^2 = .60$ . The smaller RMSE of item discrimination was associated with larger sample size ( $F(1, 464) = 150.28$ ,  $p < .001$ ,  $\eta_p^2 = .25$ ), more items in each domain ( $F(1, 464) = 16.33$ ,  $p < .001$ ,  $\eta_p^2 = .03$ ), and lower correlation levels between domains ( $F(1, 464) = 7.36$ ,  $p = .007$ ,  $\eta_p^2 = .02$ ). However, the number of domains had no significant effect on the RMSE of item discrimination estimates,  $F(1, 464) = 1.04$ ,  $p = .308$ ,  $\eta_p^2 = .002$ . Table 4.1 reports the estimated mean and standard error of RMSEs for each level of the between-subject IVs.

Table 4.1 Mean and SE of Item Discrimination RMSE by Between-subjects IVs

Between-subjects IV	Low-level		High-level	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Number of domains (D)	0.156	0.002	0.153	0.002
Number of examinees (N)	0.173	0.002	0.136	0.002
Number of items (J)	0.160	0.002	0.148	0.002
Correlation levels ( $\gamma$ )	0.150	0.002	0.158	0.002

In addition, the effect of the number of items on the RMSE of item discrimination estimates was significantly different between the two estimation methods,  $F(1, 464) = 33.30, p < .001, \eta_p^2 = .067$ . In order to further examine this pattern, a simple main effect of number of items was performed at each estimation method. As shown in Figure 4.1, for the WLSMV estimation method, larger number of items in each domain yielded smaller RMSE ( $M_{I=5} = 0.175; M_{I=10} = 0.158$ ),  $F(1, 478) = 23.16, p < .001, \eta_p^2 = .05$ . However, for the MCMC estimation method, there was no significant difference between the two levels of number of items ( $M_{I=5} = 0.145; M_{I=10} = 0.139$ ),  $F(1, 478) = 3.76, p = .053, \eta_p^2 = .01$ . This indicates that the WLSMV method seems to be more likely to be affected by the number of items compared to the MCMC method. Table 4.2 reports the RMSE and bias of item discrimination estimates for all the conditions. The negative bias under all the conditions means that the discrimination parameters are likely to be underestimated by both methods.

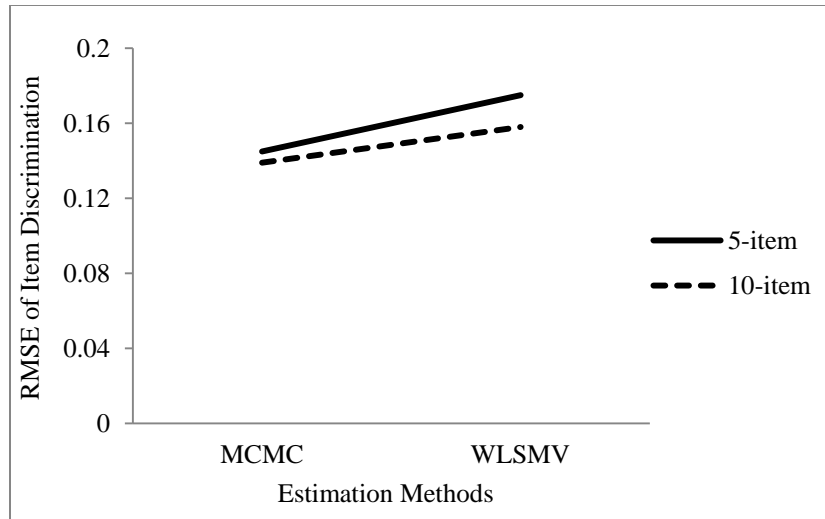


Figure 4.1 RMSE of Item discriminations by estimation methods and number of items

Table 4.2 RMSE and BIAS of Item Discrimination Estimated by MCMC vs. WLSMV

N	J	$\gamma$	D=3		D=5		
			MCMC	WLSMV	MCMC	WLSMV	
RMSE							
500	5	0.4~0.6	0.16	0.18	0.16	0.19	
		0.6~0.8	0.17	0.20	0.16	0.21	
	10	0.4~0.6	0.17	0.17	0.15	0.17	
		0.6~0.8	0.16	0.18	0.15	0.18	
1000	5	0.4~0.6	0.13	0.16	0.12	0.15	
		0.6~0.8	0.14	0.16	0.13	0.15	
	10	0.4~0.6	0.12	0.14	0.11	0.13	
		0.6~0.8	0.12	0.14	0.13	0.15	
BIAS							
500	5	0.4~0.6	-0.05	-0.02	-0.02	-0.01	
		0.6~0.8	-0.07	-0.04	-0.06	-0.02	
	10	0.4~0.6	-0.08	-0.04	-0.05	-0.02	
		0.6~0.8	-0.07	-0.01	-0.07	-0.02	
1000	5	0.4~0.6	-0.05	-0.07	-0.04	-0.05	
		0.6~0.8	-0.05	-0.06	-0.03	-0.03	
	10	0.4~0.6	-0.05	-0.05	-0.03	-0.03	
		0.6~0.8	-0.05	-0.03	-0.05	-0.04	

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

Similarly, for the item threshold estimates, a mixed ANOVA was performed on the RMSE as a function of all the conditions examined in the current study. The results showed that the RMSE obtained by MCMC ( $M = 0.109$ ,  $SE = 0.001$ ) was slightly smaller than those obtained by WLSMV ( $M = 0.113$ ,  $SE = 0.001$ ),  $F(1, 464) = 38.91$ ,  $p < .001$ ,  $\eta_p^2 = .08$ . The RMSE that resulted from a sample size of 1000 ( $M = 0.095$ ,  $SE = .001$ ) was significant smaller than that from a sample size of 500 ( $M = 0.128$ ,  $SE = 0.001$ ),  $F(1, 464) = 330.38$ ,  $p < .001$ ,  $\eta_p^2 = .42$ . The samples with higher correlations between domains ( $M = 0.115$ ,  $SE = 0.001$ ) yielded larger RMSE than the samples with lower correlations between domains ( $M = 0.108$ ,  $SE = 0.001$ ),  $F(1, 464) = 13.01$ ,  $p < .001$ ,  $\eta_p^2 = .03$ , but the magnitude of this effect was quite small. Number of items and number of domains had no significant effect on the RMSE of item threshold estimates ( $F(1, 464) = 2.19$ ,  $p = .139$ ,  $\eta_p^2 = .01$ ;  $F(1, 464) = 3.52$ ,  $p = .061$ ,  $\eta_p^2 = .01$ ). All the interaction effects were either not statistically significant or accounted for very small amount of variance (i.e.,  $\eta_p^2 < .03$ ) in the RMSE of item threshold estimates. Table 4.3 shows the RMSE and bias of item threshold estimates by the two estimation methods for all the conditions. For both estimation methods, both RMSE and bias of threshold estimates were smaller than those of discrimination estimates (Table 4.2), indicating the threshold estimates were better recovered than the discrimination estimates.

Table 4.3 RMSE and BIAS of Item Threshold Parameters Estimated by MCMC vs. WLSMV

N	J	$\gamma$	D=3		D=5	
			MCMC	WLSMV	MCMC	WLSMV
RMSE						
500	5	0.4~0.6	0.12	0.13	0.12	0.13
		0.6~0.8	0.13	0.14	0.13	0.14
	10	0.4~0.6	0.13	0.12	0.12	0.13
		0.6~0.8	0.13	0.12	0.13	0.13
1000	5	0.4~0.6	0.09	0.09	0.09	0.10
		0.6~0.8	0.10	0.10	0.10	0.10
	10	0.4~0.6	0.09	0.09	0.09	0.09
		0.6~0.8	0.09	0.10	0.10	0.10
BIAS						
500	5	0.4~0.6	-0.01	-0.01	0.00	0.00
		0.6~0.8	-0.01	0.00	0.00	0.00
	10	0.4~0.6	0.00	0.01	-0.01	0.00
		0.6~0.8	-0.01	0.00	0.00	0.00
1000	5	0.4~0.6	-0.01	-0.01	0.00	0.00
		0.6~0.8	0.00	0.00	0.00	0.01
	10	0.4~0.6	-0.01	0.00	0.00	0.00
		0.6~0.8	0.00	0.00	0.00	0.01

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

## 4.2 RECOVERY OF REGRESSION COEFFICIENTS

To examine the accuracy of the recovery of regression coefficients ( $\gamma$ ) of domain abilities on the overall ability for both HO-IRT and second-order CFA models, the RMSEs of  $\gamma$  were computed, and a mixed ANOVA was performed on the RMSE as a function of estimation methods, number of domains, number of examinees, number of items in each domain, and correlation between domains.

Overall, the MCMC method ( $M = 0.05$ ,  $SE = 0.001$ ) produced slightly smaller RMSE of regression coefficients than the WLSMV method ( $M = 0.06$ ,  $SE = 0.001$ ),  $F(1, 464) = 57.59$ ,  $p < .001$ ,  $\eta_p^2 = .11$ . The RMSE was significantly smaller when the number of domains was higher

( $F(1, 464) = 16.41, p < .001, \eta_p^2 = .03$ ), the sample size was larger ( $F(1, 464) = 35.12, p < .001, \eta_p^2 = .07$ ), the number of items was larger ( $F(1, 464) = 11.81, p = .001, \eta_p^2 = .03$ ), and the correlation between domains was higher ( $F(1, 464) = 13.85, p < .001, \eta_p^2 = .03$ ). Table 4.4 reports the mean and standard error of RMSEs by the between-subjects IVs with two specified levels. As indicated by Table 4.4, the mean differences between the high and low levels of all the between-subjects IVs were quite small (no more than 0.008) except for the sample size. Although these small differences were statistically significant, their effect sizes (about .03) were quite small as seen in the ANOVA analyses.

Table 4.4 Mean and SE of RMSE of Regression Coefficients by Between-subjects IVs

Between-subjects IV	Low-level		High-level	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Number of domains (D)	0.057	0.001	0.049	0.001
Number of examinees (N)	0.059	0.001	0.048	0.001
Number of items (J)	0.057	0.001	0.050	0.001
Correlation levels ( $\gamma$ )	0.057	0.001	0.050	0.001

As indicated by Figure 4.2, the difference on the RMSE between the two estimation methods was dependent on the degree of correlation among domains,  $F(1, 464) = 132.96, p < .001, \eta_p^2 = .22$ . When the regression coefficients were small, which indicates low correlation between domains, the MCMC method ( $M = 0.049, SE = 0.001$ ) produced smaller RMSE than the WLSMV method ( $M = 0.065, SE = 0.002$ ),  $F(1, 239) = 103.57, p < .001, \eta_p^2 = .30$ . On the other hand, when the regression coefficients were large, the MCMC method ( $M = 0.051, SE = 0.001$ ) resulted in larger RMSE than the WLSMV method ( $M = 0.048, SE = 0.001$ ),  $F(1, 239) = 16.29, p < .001, \eta_p^2 = .06$ . However, the impact of correlation between domains seems to be stronger on

the WLSMV method. The RMSEs of regression coefficients under all the conditions were shown in Table 4.5.

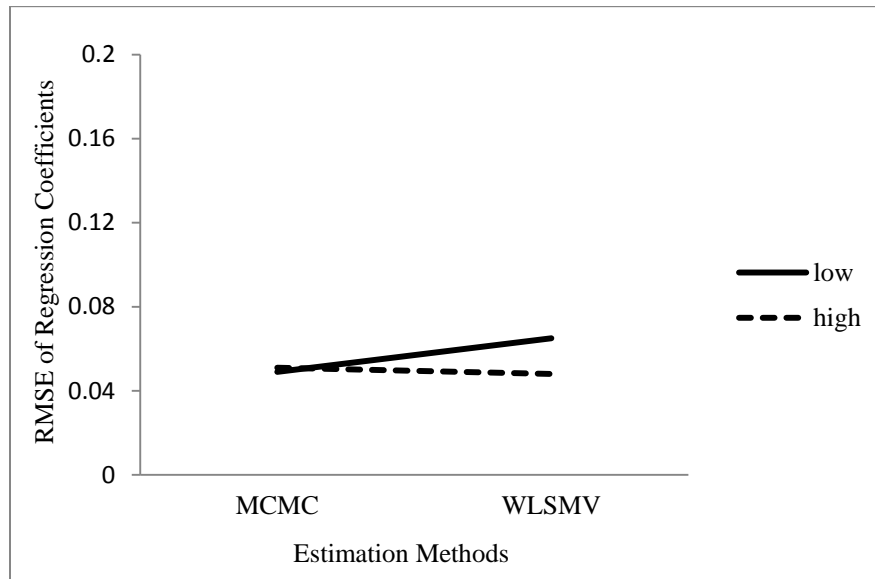


Figure 4.2 RMSE of regression coefficients by estimation method and correlation levels between domains

Table 4.5 RMSE of Regression Coefficients Estimated by MCMC vs. WLSMV

N	J	$\gamma$	D=3		D=5	
			MCMC	WLSMV	MCMC	WLSMV
RMSE						
500	5	0.4~0.6	0.06	0.10	0.05	0.07
		0.6~0.8	0.06	0.06	0.06	0.05
	10	0.4~0.6	0.05	0.07	0.05	0.06
		0.6~0.8	0.06	0.05	0.06	0.05
1000	5	0.4~0.6	0.06	0.07	0.04	0.05
		0.6~0.8	0.05	0.04	0.05	0.04
	10	0.4~0.6	0.05	0.06	0.04	0.04
		0.6~0.8	0.05	0.04	0.04	0.04

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

### 4.3 ABILITY RECOVERY

The recovery of ability estimates for the two estimation methods was evaluated by examining the correlation between the estimated and true ability at both overall and domain levels. The higher the degree of correlation, the better recovery of the ability estimates. To compare the recovery of overall ability between the HO-IRT model and second-order CFA model, a mixed ANOVA was performed on the correlation between the estimated and true overall ability as a function of estimation methods, number of domains, number of examinees, number of items in each domain, and correlation between domains. The results showed that the correlation obtained by MCMC ( $M = .777$ ,  $SE = 0.001$ ) was a slightly higher than the correlation obtained by WLSMV ( $M = .775$ ,  $SE = 0.001$ ),  $F(1, 464) = 86.12$ ,  $p < .001$ ,  $\eta_p^2 = .16$ . However, this difference (.002) was too small to have a practical significance in terms of the correlation measure, which indicates that both estimation methods performed similarly in recovering the overall ability.

For both estimation methods, the overall ability was better recovered as the number of domains increased ( $F(1, 464) = 685.63$ ,  $p < .001$ ,  $\eta_p^2 = .60$ ), as the number of items in each domain increased ( $F(1, 464) = 58.43$ ,  $p < .001$ ,  $\eta_p^2 = .11$ ), and as the correlation between domains (i.e.,  $\gamma$ ) increased ( $F(1, 464) = 2851.69$ ,  $p < .001$ ,  $\eta_p^2 = .86$ ). The mean and standard error of the correlations between the estimated and true overall ability for all the between-subjects IVs were reported in Table 4.6. However, unlike its large impact on the item parameter estimates, sample size had no significant effect on the overall ability estimates for both estimation methods ( $F(1, 464) = 0.06$ ,  $p = .806$ ,  $\eta_p^2 = .00$ ). As shown in Table 4.6, the correlation values produced by the two levels of sample size were the same. In addition, for both estimation methods, the effect of number of domains was dependent on the correlation between domains,  $F(1, 464) = 29.66$ ,  $p <$



.001,  $\eta_p^2 = .06$ . When the correlation between domains was low, the overall ability was recovered better for the higher number of domains ( $M_{D=3} = .66$ ;  $M_{D=5} = .75$ );  $F(1, 476) = 443.71$ ,  $p < .001$ ,  $\eta_p^2 = .48$ . When the correlation between domains was high, better recovery of the overall ability was also associated with the larger number of domains ( $M_{D=3} = .82$ ;  $M_{D=5} = .88$ ),  $F(1, 476) = 190.74$ ,  $p < .001$ ,  $\eta_p^2 = .29$ . Nonetheless, the effect of increasing the number of domains on improving the overall ability estimates was a bit stronger when the correlation between domains was low. Table 4.7 presents the correlations between the estimated and true overall ability obtained from two estimation methods for all the conditions.

Table 4.6 Mean and SE of Correlation between Estimated and True Overall Ability by Between-subjects IVs

Between-subjects IV	Low-level		High-level	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Number of domains (D)	0.74	0.002	0.81	0.002
Number of examinees (N)	0.78	0.002	0.78	0.002
Number of items (J)	0.77	0.002	0.79	0.001
Correlation levels ( $\gamma$ )	0.79	0.002	0.85	0.002

Table 4.7 Correlation between True and Estimated Overall Ability

N	J	$\gamma$	D=3		D=5	
			MCMC	WLSMV	MCMC	WLSMV
RMSE						
500	5	0.4~0.6	0.64	0.63	0.75	0.74
		0.6~0.8	0.81	0.81	0.87	0.87
	10	0.4~0.6	0.68	0.67	0.76	0.75
		0.6~0.8	0.83	0.83	0.89	0.89
1000	5	0.4~0.6	0.66	0.66	0.74	0.74
		0.6~0.8	0.81	0.81	0.86	0.86
	10	0.4~0.6	0.67	0.67	0.75	0.75
		0.6~0.8	0.83	0.83	0.89	0.89

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

With respect to domain ability estimation, Tables 4.8 and 4.9 report the correlation between the true and estimated domain abilities for both estimation methods under all the conditions. As found for the overall ability, both estimation methods performed similarly in recovering the domain abilities. The higher correlation values compared to those for overall ability estimates in Table 4.7 indicate both methods performed better in recovering the domain abilities than in the recovery of the overall ability. This might be due to the smaller number of domains (i.e., 3 or 5) involved in estimating the overall ability compared to the larger number of items (i.e., 5 or 10) for estimating each domain ability. For both estimation methods, the recovery of domain abilities improved with higher numbers of items and higher correlations between domains. To examine whether increasing the number of domains improve domain ability estimates, the ability estimates of the first three dimensions under the five-dimension condition were compared to those under the three-dimension condition. It appears that increasing the number of domains had no noticeable impact on the domain ability estimates for both estimation methods. Finally, sample size showed no evident impact on the domain ability estimates.

Table 4.8 Correlation between True and Estimated Domain Abilities for D=3

N	J	$\gamma$	MCMC			WLSMV		
			$\theta_1$	$\theta_2$	$\theta_3$	$\theta_1$	$\theta_2$	$\theta_3$
500	5	0.4~0.6	0.84	0.90	0.93	0.84	0.90	0.93
		0.6~0.8	0.85	0.90	0.93	0.85	0.90	0.93
	10	0.4~0.6	0.90	0.94	0.96	0.91	0.95	0.96
		0.6~0.8	0.91	0.95	0.96	0.91	0.95	0.96
1000	5	0.4~0.6	0.83	0.90	0.93	0.83	0.90	0.93
		0.6~0.8	0.85	0.90	0.93	0.85	0.90	0.93
	10	0.4~0.6	0.90	0.94	0.96	0.91	0.94	0.96
		0.6~0.8	0.91	0.95	0.96	0.91	0.95	0.96

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

Table 4.9 Correlation between True and Estimated Domain Abilities for D=5

N	J	$\gamma$	MCMC					WLSMV				
			$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$
500	5		0.83	0.90	0.93	0.83	0.90	0.83	0.90	0.93	0.83	0.89
			0.86	0.91	0.94	0.86	0.91	0.86	0.91	0.93	0.86	0.91
	10		0.90	0.94	0.96	0.91	0.94	0.90	0.94	0.96	0.90	0.94
			0.91	0.95	0.96	0.91	0.95	0.91	0.95	0.96	0.91	0.95
1000	5		0.84	0.90	0.93	0.84	0.90	0.84	0.90	0.93	0.83	0.90
			0.85	0.90	0.93	0.85	0.90	0.85	0.90	0.93	0.85	0.90
	10		0.91	0.95	0.96	0.90	0.95	0.91	0.94	0.96	0.90	0.94
			0.91	0.95	0.96	0.91	0.94	0.91	0.95	0.96	0.91	0.94

Note: N = number of examinees; J = number of items in each domain;  $\gamma$  = regression coefficients of domain abilities on the overall ability (correlation between domains); D = number of domains.

#### 4.4 REAL DATA APPLICATION

To further investigate the behavior of the two estimation methods used in the simulation study, they were applied to a real dataset from the Beck Depression Inventory (BDI, Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) administered to 1096 Swedish nonclinical adolescents aged from 13 to 18 years (Byrne & Campbell, 1999). The BDI includes 21 four-point scale items that measure the severity of depression. Research on the factorial structure of the BDI has indicated that the inventory is most appropriately represented by a higher-order structure including three lower-order factors, namely, Negative Attitude (10 items), Performance Difficulty (7 items), and Somatic Elements (4 items), and one second-order factor (General Depression) (Byrne & Baron, 1993; Byrne, Baron, & Balev, 1998; Byrne, Baron, Larsson, & Melin, 1995). Similar to the simulation study, the real data were analyzed by a HO-IRT model with MCMC and a second-order CFA model with WLSMV, respectively.

The results of the WLSMV estimation method in Mplus showed that the second-order CFA model yielded negative residual variance related to the second factor (Performance

Difficulty), which might be due to its high correlation with other factors. Therefore, the regression coefficient of this factor on the general factor was constrained to 1 in order to properly identify the model. This constrained model in general fits the data well (RMSEA = 0.041, 90% CI = [0.036, 0.045]; CFI = 0.963, TLI = 0.958). The MCMC estimation employed for the real data was similar to the preliminary analysis in the simulation study. Three chains with randomly generated initials were run in WinBUGS, each with 5,000 iterations as burn-in, and a total of 15,000 iterations. Each chain was thinned by taking every 10<sup>th</sup> iteration to reduce the autocorrelation. To determine the model convergence with MCMC, a visual inspection of the history plots, BGR plots, and autocorrelation plots was conducted. These plots indicated that all the model parameters converged except for the regression coefficients of the second factor, i.e., Performance Difficulty ( $\gamma_2$ ). The parameter estimates were obtained based on the draws of all three chains.

In addition, several sets of Bayesian priors for the item parameters were compared for the MCMC estimation in WinBUGS. The priors specified in the simulated data were quite informative; i.e., the prior distributions of the item parameters were entirely consistent with the distributions from which the item parameters were generated. This might account for the higher degree of parameter recovery for the MCMC estimation method observed in the simulated data. With the real data analysis, therefore, the MCMC was implemented with different sets of less informative priors. First, priors similar to those used in the simulated data were specified. The discrimination parameter was estimated by setting the prior distribution as a uniform distribution with values ranging between 1 to 2.5, i.e.,  $a \sim U(1, 2.5)$ . The prior distributions for the three threshold parameters uniformly lay between -3 and 3, i.e.,  $b_1 \sim U(-3, 3)$ ,  $b_2 \sim U(b_1, 3)$ , and  $b_3 \sim U(b_2, 3)$ . Another set of prior distributions for the item parameters was also specified as uniform but

with a wider range or less informative, i.e.,  $a \sim U(1, 2.5)$ ,  $b_1 \sim U(-4, 4)$ ,  $b_2 \sim U(b_1, 4)$ , and  $b_3 \sim U(b_2, 4)$ . Finally, other commonly used priors for IRT item parameters were considered. For example, it is commonly assumed for IRT models that the log of the discrimination parameter has a normal distribution, and the location parameter is normally distributed (Hartwell & Baker, 1991). Therefore, the priors were specified for the data as  $a \sim \text{LogN}(0, 1)$ ,  $b_{1, 2, 3} \sim N(0, 4)$  (precision is 0.25 in WinBUGS) where  $b_1 < b_2 < b_3$ .

As in the simulation study, the parameters estimated from the second-order CFA were converted to the corresponding IRT parameters for the purpose of comparison. Table 4.10 shows the item parameters estimated by WLSMV and MCMC with different item priors for each BDI subscale averaged across the items. Generally speaking, the item discrimination parameter estimates obtained by MCMC were slightly higher while the threshold parameter estimates were slightly smaller compared to the corresponding estimates from WLSMV. This difference was particularly evident when the smaller range of uniform priors was used with MCMC, indicating this set of priors might poorly match to the estimated parameters. A larger variance in the prior distributions of the threshold parameters might be expected as the BDI inventory probably has higher location values to measure severe depression. Meanwhile, within the MCMC estimation method, some variability in the item parameter estimates was observed across different priors. Appendix D provides the parameter estimates for the 10 items that measure the Negative Attitude using both WLSMV and MCMC with different priors. These results indicate that the accuracy of item parameter estimates in MCMC was affected by the specified prior distributions, and MCMC performed no better than WLSMV based on the priors used for the real data.

Table 4.10 Mean Item Parameter Estimates for the BDI Subscales

Estimation Methods (Priors)	Subscales											
	Negative Attitude				Performance Difficulty				Somatic Elements			
	a	b <sub>1</sub>	b <sub>2</sub>	b <sub>3</sub>	a	b <sub>1</sub>	b <sub>2</sub>	b <sub>3</sub>	a	b <sub>1</sub>	b <sub>2</sub>	b <sub>3</sub>
WLSMV	2.05	1.17	2.06	2.92	1.52	1.00	2.72	3.51	1.29	1.69	3.26	4.03
MCMC	2.13	1.06	1.94	2.74	1.87	0.78	2.26	2.86	1.90	1.15	2.34	2.88
a~ $U(1, 2.5)$ b~ $U(-3, 3)$												
MCMC	2.00	1.21	2.18	3.15	1.61	0.85	2.67	3.50	1.55	1.44	2.89	3.62
a~ $U(0.5, 2.5)$ b~ $U(-4, 4)$												
MCMC	2.26	1.11	1.99	2.88	1.67	0.91	2.59	3.49	1.49	1.49	3.01	3.90
a~ $\log N(0,1)$ b~ $N(0, 4)$												

The relationship between the overall depression and each of the subscales estimated by MCMC were almost identical as those estimated by WLSMV regardless of prior specifications (Table 4.11). This is expected as the prior distributions of the item parameters have little impact on the estimation of these regression coefficients. Although the  $\gamma_2$  estimated by MCMC did not converge based on the iterations specified in this analysis, its value was very close to 1 as constrained in the CFA model. This also indicates both estimation methods might have problems in estimating highly correlated factors.

Table 4.11 Estimated Regression Coefficients between the Overall and Subscale Scores for the BDI Data

Estimation Methods (Priors)	Subscales		
	Negative Attitude	Performance Difficulty	Somatic Elements
WLSMV	0.87	1.00*	0.85
MCMC	0.89	0.99	0.86
$a \sim U(1, 2.5)$ $b \sim U(-3, 3)$			
MCMC	0.88	0.99	0.85
$a \sim U(0.5, 2.5)$ $b \sim U(-4, 4)$			
MCMC	0.88	0.99	0.87
$a \sim \log N(0, 1)$ $b \sim N(0, 4)$			

\* constrained to be 1.

In order to compare the two methods in ability estimation, the correlation between the factor scores estimated from the second-order CFA model and ability estimates obtained from the HO-IRT model was computed for both general and domain scales. As shown in Table 4.12, the two measurement models yielded almost perfect correlation in ability estimates, indicating they were almost equivalent in ranking the examinees' performance at both general and specific levels. In conclusion, the results from the real data analysis showed the item and person parameter estimates produced by the two estimation methods (MCMC vs. WLSMV) specified under the two frameworks (IRT vs. CFA) were highly comparable.

Table 4.12 Correlation of Ability Estimates between MCMC and WLSMV for the BDI Data

Estimation Methods (Priors)	Subscales			
	Negative Attitude	Performance Difficulty	Somatic Elements	General
MCMC $a \sim U(1, 2.5)$ $b \sim U(-3, 3)$	0.997	0.996	0.993	0.996
MCMC $a \sim U(0.5, 2.5)$ $b \sim U(-4, 4)$	0.997	0.997	0.997	0.997
MCMC $a \sim \log N(0,1)$ $b \sim N(0, 4)$	0.998	0.996	0.997	0.997



## **5.0 DISCUSSION**

This project was intended to examine the accuracy of MCMC and WLSMV in recovering item and ability parameters with polytomous response data. Using a simulation approach, the study also aimed to evaluate how the estimation methods specified under the two frameworks (IRT vs. CFA) were affected by the variables considered in this study, including number of dimensions, sample size, test length in each dimension and correlation between dimensions. This section summarizes the major findings from this work and discusses some potential implications for psychometric analysis. In addition, the limitations of the present study and future research directions are provided.

### **5.1 SUMMARY AND IMPLICATIONS**

#### **5.1.1 MCMC vs. WLSMV in Parameter Estimation**

Overall, the accuracy of the MCMC and WLSMV methods in the estimation of both item and ability parameters was quite comparable. As far as the item discrimination and threshold parameter estimates are concerned, there were only very small differences in the RMSE between the two estimation methods. The MCMC estimation performed marginally better than the WLSMV method only in the discrimination parameter estimates. However, this slight superiority

for the MCMC method might be due to the strong prior specified in the simulated data, i.e., the prior perfectly matched the generating parameters. By using different sets of Bayesian priors in the real data, the results showed some variations in the item parameter estimates across different priors for the MCMC method, which further confirms the role of priors in the accuracy of item parameter estimation. This concurs with findings from Kieftenbeld and Natesan's (2012) simulation study that compared the MCMC and MML in recovering graded response model parameters. They concluded that the accuracy of item parameter estimates in MCMC would probably be improved if a prior was better matched to the specific generating parameters. In addition, for both estimation methods used in the current study, the item threshold parameters were better recovered than the discrimination parameters under all the conditions.

With regard to the overall ability estimation, both estimation methods performed similarly. As shown by both simulated and real data, the overall ability estimates produced by the two estimation methods were almost identical in terms of the degree of their correlations with the true overall ability. Although the simulation study indicates the correlation between the true and estimated overall ability was slightly higher for the MCMC method, the difference was negligible. This similarity between the two estimation methods has also been found in the estimation of domain abilities. Additionally, both estimation methods performed similarly in the recovery of regression coefficients of domain abilities by the overall ability.

### **5.1.2 Effects of Between-subjects Independent Variables**

Generally speaking, sample size, number of items, and correlation between domains were found to have an impact on the recovery of item parameter estimates for both estimation methods. Number of dimensions, however, had no significant effect on the item parameter estimation

based on the evaluation criteria used in the current study. Consistent with previous research (de la Torre & Hong, 2010; Finch, 2010; Yao, 2010; Wirth & Edwards, 2007), this study found that both estimation methods provided more accurate item parameter estimates as the sample size increased. In addition, sample size accounted for the most in the accurate recovery of item parameters compared to other between-subject IVs.

In de la Torre and Hong's (2010) simulation study involving dichotomous item calibration, the HO-IRT model produced more accurate item parameter estimates as the number of items increased. This result seems to hold for the polytomous items examined in the current study. However, this effect was more noticeable for the WLSMV method in estimating the item discrimination parameters. Increasing the number of items showed no significant difference on the threshold estimates for either estimation method.

Some previous studies indicated that MCMC or adjusted WLS increased the accuracy of item estimation as the correlation between domains increased (e.g., Finch, 2010; Yao, 2010). This, however, was not observed in the current study. On the contrary, this study found that the accuracy of item parameter estimates tended to decrease as the correlation between domains increased although this trend was not strong. In fact, the impact of correlation between domains on the item parameter estimation requires further investigation. Yao (2010) compared different estimation models, including HO-IRT and MIRT, and claimed that the RMSE for item parameters decreased as the correlation between dimensions increased. However, this pattern, in fact, only held for the discrimination parameters in her study. Furthermore, unlike most previous studies that assumed the degree of correlations between domains were equal (e.g., Finch, 2010; Yao, 2010), this study used varied correlations among the dimensions which might have led to the different conclusion. Another plausible explanation is that the correlation levels specified in

the current study were not high enough to observe the patterns indicated by the previous research. That is, the correlation between domains might have to reach a specific level to show its effect on improving the accuracy of item parameter estimates. Thus, the results of the current study, together with previous findings on both estimation methods, are not sufficient to reach a conclusion about the impact of the correlation between domains on the item parameter estimation and further investigation is required.

For the regression coefficients of domain ability predicted by the overall ability, more accurate estimates were obtained for the MCMC and WLSMV methods as the number of domains, the sample size, and the test length increased. These findings were consistent with previous studies on the HO-IRT models (de la Torre & Hong, 2010; de la Torre & Song, 2009). However, the two estimation methods performed differently depending on the degree of correlation between domains. MCMC recovered the regression coefficients better when the correlation between domains was low while WLSMV produced more accurate estimates when the correlation between domains was high. Nevertheless, WLSMV was more sensitive to the impact of correlation between domains in terms of its effect size. In practice, if the regression coefficients are of interest, MCMC method may be more appropriate.

As found for dichotomous response data (de la Torre & Hong, 2010; de la Torre & Song, 2009), the HO-IRT and CFA models with polytomous items in the current study provided more accurate overall ability estimates with the longer test, greater number of dimensions, and higher correlations between domains. Meanwhile, for both estimation methods, the accuracy of domain ability estimates improved when the number of items was higher and when the correlations between the domains were higher. However, unlike previous studies that found better domain ability estimates were associated with a larger number of dimensions, the impact of the number

of dimensions on the domain ability estimates was not clearly observed in the current study. This might be due to the varied degree of correlations between domains specified in the current study. Therefore, according to the previous findings and results of this study, the number of domains may play a role in improving the accuracy of domain ability estimation only when the correlations between domains are at least moderately correlated (about 0.5). In other words, to improve the current domain ability estimates by increasing the number of domains is only warranted when the added domains are moderately correlated with the current dimension or highly correlated with the higher-order dimension.

### **5.1.3 Implications for Practice**

One of the most relevant implications from the present study for practitioners is to decide which method to use. A general result of this study is that the behavior of MCMC and WLSMV is similar with respect to parameter estimation accuracy. With the conditions examined in the current study, there is little superiority in one method over the other. However, in terms of estimation speed, WLSMV has a clear advantage over MCMC. The much slower estimation speed for MCMC is partly due to the software used in the current study. As few software options are available for the MCMC implementation, increasing the efficiency of MCMC estimation can only be attended by writing programs tailored to a specific application. Furthermore, the convergence criteria used for MCMC are usually diverse, and most of them are based on visual inspection. With a variety of widely accepted software in SEM, on the other hand, diagnosis of model convergence and goodness of fit is much easier for WLSMV. In addition, inappropriately specified prior distributions also present a problem for parameter estimation in MCMC. Therefore, it can be argued that, in the conditions as the current study, WLSMV is preferred in

applied research due to its greater efficiency, easier diagnosis of model convergence, and no requirement for prior distributions.

Another relevant implication for applied researchers refers to the relationship between the number of dimensions and correlation between dimensions in improving domain ability estimation. The primary advantage of a multidimensional or higher-order model structure over a unidimensional model is that the accuracy of domain ability estimation can be improved by borrowing information from other domains. With that being said, the additional domains should be at least moderately correlated with the estimated domain (about 0.5 or above). This is a prerequisite for using a higher number of dimensions to improve domain ability estimation. Also, the interpretation and use of the scores obtained from such a higher-order model needs a further consideration. Although the two estimation models (HO-IRT and second-order CFA) provide almost identical score ranks at both overall and subscale levels, it is important to consider the interpretation and intended uses of these scores. Since the subscale scores are obtained based on the information borrowed from other tests or subscales, the interpretation of these scores is more complex (de la Torre & Patz, 2005; Stone et al, 2010). For example, if test users want to compare examinees' performance on a specific domain, it would be inappropriate for them to use the subscale scores obtained via the higher-order approach because they do not have a straightforward interpretation that reflects examinee's performance on that specific domain. However, if teachers intend to diagnose students' weaknesses in order to plan remedial instruction, their use of subscale scores through the approach could be supported since they are more accurate. As de la Torre and Song (2009) pointed out, the domain abilities are more appropriate for within-person comparisons whereas overall ability is more appropriate for between-subject comparisons.

## 5.2 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

This study used a simulation approach to compare MCMC and WLSMV estimation methods under higher-order IRT and CFA frameworks, respectively. Several limitations were recognized. First, similar to most simulation studies, it would be impossible to include all the possible variables or variable levels within one study. Therefore, the results may not generalize to other situations not considered in the current study. For example, in this study, the number of response category was fixed (5-category) and ability distribution was assumed to be normal. Further studies can be performed to examine how changing these conditions could impact the two estimation methods. Though the levels of variables were carefully designed in the current study, more comprehensive levels could be included to more fully explore the extent to which these variables affect the performance of the two estimation methods. Particularly, the sample size used in the current study seems to be minimally adequate for GR model estimation. Future research can investigate the performance of the two estimation methods for smaller sample sizes. This seems an important topic because sufficiently large sample size is not easily accessible especially for many psychological tests.

Secondly, the data were generated from the same model used for the parameter estimation. The higher-order structure used in the current study assumes a linear relationship between overall and specific abilities and only one higher-order factor was included. However, with real data, the correlation structure might be more complex than the one specified in the current study, such as exhibiting a nonlinear relationship between overall and specific factors and multiple higher-order factors. It would be useful to investigate the extent to which the two estimation methods can be affected by the violation of the current model structure. In other

words, an important future research question can examine how well the two estimation methods perform when the model is misspecified to some degree.

Furthermore, the higher-order models formulated in the current study assume between-item multidimensionality at the lower order. However, in actual practice it is not uncommon to see some items loading on more than one dimension, i.e., within-item multidimensionality. Therefore, incorporating such items into both higher-order IRT and CFA models and investigating their performance in parameter estimation would be a possible direction for future research. It would be also helpful to examine the degree to which the estimation methods can withstand the violation of the assumption of between-item multidimensionality. In addition, the higher-order model was mathematically related to a bi-factor model which includes a single general factor for all items and one or more orthogonal specific factors for some or all of items. Therefore, in future research, the performance of bi-factor models can be compared to the higher-order models under both IRT and CFA frameworks.

Finally, although different sets of Bayesian priors were used for the real data in the current study, it is not clear to what extent the accuracy of parameter estimation can be affected by the specification of prior distributions. As the simulation study used the same prior distributions as their simulating distributions, it would be useful to examine how the prior distributions that are deviated from the true distributions of parameters would impact the estimation results. Meanwhile, Bayesian estimation has also been developed and applied to the structural equation modeling. Further research can focus on the comparison between higher-order IRT and CFA models with both using Bayesian estimation.



## APPENDIX A

### SAS CODE USED TO GENERATE HIGHER-ORDER GRADED RESPONSE DATA

/\*\*\*\*\*\*SAS Macro used to generate item response with number of domains D = 3 and number of items in each domain J = 5\*\*\*\*\*\*/

```
%let wrkdir = C:\DS_code;  
libname mydata "&wrkdir";
```

```
%let nthres=4; /*fixed factor: number of thresholds and total number of items*/  
%let nitems=15;
```

```
%let npersons1=500; /* IV: sample size*/  
%let npersons2=1000;
```

```
%let corr1= 0.4; /*IV: correlation among subscales*/  
%let corr2= 0.6;
```

```
%let nrep =30;
```

```
%macro sim;  
%do rep=1 %to &nrep;
```

```
%do f1=1 %to 2; /*the 3rd factor */  
%if &f1=1 %then %let npersons=&npersons1;  
%else %if &f1=2 %then %let npersons=&npersons2;
```

```
%do f2=1 %to 2; /*the 4th factor */  
%if &f2=1 %then %let corr=&corr1;  
%else %if &f2=2 %then %let corr=&corr2;
```

```
%let seed = %eval(100000+10000+&f1*1000+&f2*100+&rep-1);
```

```

proc iml;
/*generate item parameters*/
/* generate thresholds*/
call randseed(&seed);

b1 = j(&nitems,1,0);
call randgen (b1, 'uniform');
b1= -1*b1-1;
b=b1||j(&nitems,3,0);

do j=1 to &nitems;
    pick =j(&nitems,1,0);
    call randgen (pick, 'uniform');

    if pick[j] <=0.5 then do;
        b[j,2] =b[j,1]+0.75;
        b[j,3]= b[j,2]+0.75;
        b[j,4]=b[j,3]+0.75;
    end;

    if pick[j]> 0.5 then do;
        b[j,2] =b[j,1]+1;
        b[j,3]= b[j,2]+1;
        b[j,4]=b[j,3]+1;
    end;
end;

/* generate slope parameters*/
a =j(5,1,0);
call randgen(a,'uniform');
a1 = 0.5*a+1;
a2 = 1.1*a+1.2;
a3 = 0.5*a+2;
a =a1//a2//a3;
/*generate ability*/
theta_overall = j(&npersons,1,0);
call randgen(theta_overall,'normal',0,1);

gamma = j(3,1,0);
call randgen(gamma,'uniform');
gamma = gamma*0.2 + &corr;
eta = j(&npersons,3,0);
call randgen(eta,'normal',0,diag(sqrt(1-gamma##2)));
theta = theta_overall * t(gamma) + eta;
/*generate item responses*/
response = j(&npersons,&nitems,1);

```

```

do i = 1 to &npersons;

/* first 5 items */
do j = 1 to 5;
    pstar = j(4,1,0);
    do thres = 1 to &nthres;
        pstar[thres] = exp(a[j]*(theta[i,1] - b[j,thres]))/(1+exp(a[j]*(theta[i,1] - b[j,thres])));
    end;
    p = j(5,1,0);
    p[1] = 1 - pstar[1];
    p[2] = pstar[1] - pstar[2];
    p[3] = pstar[2] - pstar[3];
    p[4] = pstar[3] - pstar[4];
    p[5] = pstar[4];
    pcum = j(5,1,1);
    pcum[1] = p[1];
    pcum[2] = sum(p[1:2]);
    pcum[3] = sum(p[1:3]);
    pcum[4] = sum(p[1:4]);
    call randgen(r, 'uniform');
    do k = 1 to &nthres;
        if (r > pcum[k] & r <= pcum[k+1]) then response[i,j] = k + 1;
    end;
end;

/* second 5 items */
do j = 6 to 10;
    pstar = j(4,1,0);
    do thres = 1 to &nthres;
        pstar[thres] = exp(a[j]*(theta[i,2] - b[j,thres]))/(1+exp(a[j]*(theta[i,2] - b[j,thres])));
    end;
    p = j(5,1,0);
    p[1] = 1 - pstar[1];
    p[2] = pstar[1] - pstar[2];
    p[3] = pstar[2] - pstar[3];
    p[4] = pstar[3] - pstar[4];
    p[5] = pstar[4];
    pcum = j(5,1,1);
    pcum[1] = p[1];
    pcum[2] = sum(p[1:2]);
    pcum[3] = sum(p[1:3]);
    pcum[4] = sum(p[1:4]);
    call randgen(r, 'uniform');
    do k = 1 to &nthres;
        if (r > pcum[k] & r <= pcum[k+1]) then response[i,j] = k + 1;
    end;
end;
end;

```

```

/*third 5 items*/
do j = 11 to &nitems;
    pstar = j(4,1,0);
    do thres = 1 to &nthres;
        pstar[thres] = exp(a[j]*(theta[i,3] - b[j,thres]))/(1+exp(a[j]*(theta[i,3] - b[j,thres])));
    end;
    p = j(5,1,0);
    p[1] = 1 - pstar[1];
    p[2] = pstar[1] - pstar[2];
    p[3] = pstar[2] - pstar[3];
    p[4] = pstar[3] - pstar[4];
    p[5] = pstar[4];
    pcum = j(5,1,1);
    pcum[1] = p[1];
    pcum[2] = sum(p[1:2]);
    pcum[3] = sum(p[1:3]);
    pcum[4] = sum(p[1:4]);
    call randgen(r, 'uniform');
    do k = 1 to &nthres;
        if (r > pcum[k] & r <= pcum[k+1]) then response[i,j] = k + 1;
    end;
end;
end;
/*save thetas, a, b, gamma & responses*/;
overall = theta_overall||theta;
create theta_true from overall [colname={theta_overall theta1 theta2 theta3}];
append from overall;
close theta_true;

create b_true from b [colname={b_t1 b_t2 b_t3 b_t4}];
append from b;
close b_true;

create a_true from a [colname={a_t}];
append from a;
close a_true;

create gamma_true from gamma [colname={gamma_t}];
append from gamma;
close gamma_true;

create response from response;
append from response;
close response;
quit;

```

## APPENDIX B

### WINBUGS CODE FOR ESTIMATING A HO-IRT MODEL (D = 3 & J = 5)

```
model
{
for (j in 1:N) {
  for (i in 1:5) {
    for (k in 1:4) {
      logit(pstar[j, i, k]) <- a[i]*(theta[j, 1]- b[i, k])} }

  for (i in 6:10) {
    for (k in 1:4) {
      logit(pstar[j, i, k]) <- a[i]*(theta[j, 2]- b[i, k])} }

  for (i in 11:T) {
    for (k in 1: 4) {
      logit(pstar[j, i, k]) <- a[i]*(theta[j, 3]- b[i, k])} }

for (i in 1:T){
  p[j, i, 1] <- 1-pstar[j, i, 1]
  for(k in 2:4) {
    p[j, i, k] <- pstar[j, i, k-1] - pstar[j, i, k]}
  p[j, i, 5] <- pstar[j, i, 4]
  resp[j, i] ~ dcat(p[j, i, 1:5])}
}

for (j in 1:N){
  for (d in 1:3){
    theta[j, d] <- gamma[d] * theta_overall[j] + eta[j,d]
    eta[j,d] ~dnorm(mu[d], tau[d])
  }
  theta_overall[j] ~dnorm(0, 1)
}
```

```

for (d in 1:3) {
  mu[d]<-0
  tau[d]<-1/(1-pow(gamma[d],2))
  gamma[d] ~dunif(0,1)
}

```

#Priors for item parameters

```

for (i in 1:T) {
  a[i] ~ dunif(1, 2.5)
}
for (i in 1:T){
  b[i,1] ~dunif (-3,-1)
  b[i,2] ~dunif(b[i,1],3)
  b[i,3] ~dunif(b[i,2],3)
  b[i,4] ~ dunif(b[i,3],3)
}
}

```

## APPENDIX C

### MPLUS CODE FOR ESTIMATING A SECOND-ORDER CFA MODEL ( $D = 3$ & $J = 5$ )

```
TITLE: 3-factor 5-item CFA analysis with GRM;
DATA: FILE IS "C:\DS_code\Mplus\data_mplus_3by5.dat";
VARIABLE:
NAMES are i1-i15;
CATEGORICAL are i1-i15;
ANALYSIS: ESTIMATOR is WLSMV;
MODEL:
factor1 BY i1-i5;
factor2 BY i6-i10;
factor3 BY i11-i15;
gfactor BY factor1-factor3;
OUTPUT:
stdyx;
tech1;
SAVEDATA:
File is fscores_3by5.dat;
SAVE = fscores;
RESULTS are res.dat;
```

## APPENDIX D

### ITEM PARAMETER ESTIMATES FOR THE NEGATIVE ATTITUDE SUBSCALE IN BDI-II USING WLSMV AND MCMC WITH DIFFERENT BAYESIAN PRIORS

Items	WLSMV				MCMC a~ $U(1, 2.5)$ b~ $U(-3, 3)$				MCMC a~ $U(0.5, 2.5)$ b~ $U(-4, 4)$				MCMC a~ $\log N(0,1)$ b~ $N(0, 4)$			
	$a$	$b_1$	$b_2$	$b_3$	$a$	$b_1$	$b_2$	$b_3$	$a$	$b_1$	$b_2$	$b_3$	$a$	$b_1$	$b_2$	$b_3$
1	2.27	1.11	2.22	3.47	2.41	0.99	2.03	2.94	2.25	1.14	2.31	3.72	2.39	1.06	2.15	3.58
2	1.95	1.34	1.92	3.04	2.21	1.20	1.77	2.82	1.99	1.39	2.02	3.25	2.13	1.29	1.87	3.00
3	2.42	1.04	1.50	2.93	2.42	0.97	1.43	2.83	2.35	1.10	1.59	3.16	2.65	1.01	1.45	2.85
4	2.79	1.35	2.05	2.68	2.44	1.31	2.07	2.78	2.42	1.45	2.26	3.01	3.18	1.28	1.96	2.59
5	1.59	1.29	2.76	3.32	1.90	1.08	2.41	2.90	1.61	1.29	2.88	3.58	1.73	1.20	2.66	3.32
6	2.37	0.84	2.55	2.82	2.41	0.75	2.47	2.79	2.36	0.86	2.68	3.04	2.71	0.79	2.42	2.74
7	1.50	0.24	1.07	3.26	1.71	0.17	0.92	2.89	1.50	0.27	1.11	3.49	1.58	0.23	1.04	3.27
8	2.50	1.27	2.23	2.67	2.42	1.19	2.17	2.68	2.38	1.33	2.36	2.90	2.84	1.20	2.10	2.57
9	1.55	1.57	2.47	2.65	1.59	1.48	2.46	2.66	1.48	1.67	2.72	2.95	1.60	1.53	2.49	2.70
10	1.58	1.61	1.83	2.31	1.79	1.44	1.66	2.14	1.65	1.63	1.87	2.39	1.76	1.51	1.73	2.21



## BIBLIOGRAPHY

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17*, 251-269.
- Allen, D. D., & Wilson, M. (2006). Introducing multidimensional item response modeling in health behavior and health education research. *Health Education Research, 21*, i73-i84.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.
- Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Taylor & Francis.
- Beauducel, A., & Herzberg P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186-203.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.
- Bentler, P. M. (1995). *EQS Program Manual*. Encino, CA: Multivariate Software Inc.

- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541–561.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-458.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.
- Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2002). *TESTFACT 4*. Chicago: Scientific Software International.
- Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27, 395-414.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153-168.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.
- Brooks, S. P., & Roberts, G. O. (1998). Convergence assessments algorithms. *Statistics and Computing*, 8, 319–335.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.
- Byrne, B. M., & Baron, P. (1993). The Beck Depression Inventory: Testing and cross-validating a hierarchical factor structure for nonclinical adolescents. *Measurement and Evaluation in Counseling and Development*, 26, 164-178.
- Byrne, B. M., Baron, P., & Balev, J. (1998). The Beck Depression Inventory: A cross-validated test of second-order factorial structure for Bulgarian adolescents. *Educational and Psychological Measurement*, 58, 241-251.
- Byrne, B. M., Baron, P., Larsson, B., & Melin, L. (1995). The Beck Depression Inventory: Testing and cross-validating a second-order factorial structure for Swedish nonclinical adolescents. *Behavioral Research and Therapy*, 33, 345-356.

- Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-cultural Psychology*, 30, 555-574.
- Chib, S., & Greenberg, E. (1995). Understanding the Metropolis Hastings algorithm. *American Statistical Journal*, 49, 327-335.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Cohen, A. S., & Kim, S. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement*, 22, 116-130.
- Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician*, 58, 330-336.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, 20, 405-416.
- de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement*, 32, 355-370.
- de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, 33, 465-485.
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, 34, 267-285.
- de la Torre, J., & Patz, R. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30, 295-311.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher order IRT model approach. *Applied Psychological Measurement*, 33, 620-639.
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47, 309-326.
- Edwards, M. C. (2005). *MultiNorm: Multidimensional Normal Ogive Item Response Theory Analysis* [Computer software].

- Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, 75, 474-497.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14, 275-299.
- Finch, H. (2010). Item parameter estimation for MIRT model: Bias and precision of confirmatory factor analysis based models. *Applied Psychological Measurement*, 34, 10-26.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.
- Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-389.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271-288.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 45, 457-511.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004), *Bayesian data analysis* (2<sup>nd</sup> Ed). New York: Chapman & Hall.
- Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 47, 473-511.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1-16). New York: Chapman & Hall.
- Glöckner-Rist, A., & Hoijsink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10, 544-565.
- Gosz, J. K., & Walker, C. M. (2002). *An empirical comparison of multidimensional item response theory data using TESTFACT and NOHARM*. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

- Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, 35, 57-63.
- Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, 15, 375-389.
- Hill, C. D., Edwards, M. C., Thissen D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni J. W. (2007). Practical issues in the application of item response theory: A demonstration using item form the Pediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core Scales. *Medical Care*. 45, S39-S47.
- Hong, Y., Lam, T. C., & de la Torre, J. (2010). *Ancillary variables and multidimensional scoring of polytomous responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.
- Jöreskog, K., & Sörbom, D. (1996). LISREL 8 User's Reference Guide. Chicago: Scientific Software International.
- Kieftenbeld, V. & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, 36, 399-418.
- Kim, J., & Bolt, D. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26, 38-51.
- Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, 25, 163-176.
- Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26, 25-41.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2<sup>nd</sup> Ed). New York: Guilford Press.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26, 457-477.
- Lane, S. & Stone, C.A. (2006). Performance Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.
- Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, 29, 3-25.

- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- McDonald, R. P. (1997). Normal-Ogive multidimensional model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer.
- McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 187-216). Mahwah, NJ: Lawrence Erlbaum.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.
- Muthén, L. K. & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.
- Patz, R., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146-178.
- Patz, R., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366.
- Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, 46, 273-286.
- Raftery, A. E., & Lewis, S.M. (1992). One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7, 493-497.
- Reckase, M. D. (1997a). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Reckase, M. D. (1997b). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Personality and Social Psychology*, 65, 133-144.
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27-48.
- Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 45-54). London: Chapman and Hall.
- Roussos, L., & Stout, W. F. (1996). A multidimensional-based DIF analysis paradigm. *Applied Psychological Measurement*, 20, 335-371.
- Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Structural Equation Modeling*, 11, 424-451.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17*.
- Satorra, A., & Bentler, P. M. (1994). Corrections for test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.
- Segall, D. O. (2000). Principles of Multidimensional Adaptive Testing. W. J. van der Linden, and C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 53-57). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, 29, 461-488.
- Sinharay, S. (2010). How often do subscores have added value: Results from operational and simulated data. *Journal of Educational Measurement*, 47, 150-174.
- Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67, 899-919.
- Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, 68, 413-430.
- Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WINBUGS Version 1.4 User's Manual* [Computer software manual] Cambridge, UK: MRC Biostatistics Unit.

- Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, 23, 63-86.
- Stone, C. A., & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement*, 66, 193-214.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159-203.
- Thissen, D. (1991). *MULTILOG (Version 6)* [Computer software and manual]. Chicago: Scientific Software.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567- 577.
- Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116-136.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement*, 12, 239-252.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.
- Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, 26, 339-352.
- Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.
- Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, 47, 339-360.
- Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31, 83-105.



- Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, 46, 177-197.
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30, 469-492.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4<sup>th</sup> ed.), (pp. 111-154). Westport, CT: American Council on Education/Praeger.
- Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 113-128.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG (Version 3)* [Computer manual]. Chicago: Scientific Software.